Audiovisual Interaction

Influence of visual appearance on loudspeaker sound quality evaluation

Karandreas, Theodoros-Alexandros

Publication date:
2011

Document Version
Early version, also known as pre-print

Audiovisual Interaction

Influence of visual appearance on loudspeaker sound quality evaluation

Ph.D. thesis by
Alex Karandreas

Section of Acoustics
Department of Electronic Systems
Aalborg University
Aalborg, Denmark 2010

Audiovisual Interaction

Influence of visual appearance on loudspeaker sound quality evaluation

Ph.D. thesis

Alex Karandreas

November 2010

Section of Acoustics
Department of Electronic Systems
Aalborg University
Aalborg, Denmark 2010

Audiovisual Interaction
- Influence of visual appearance
on loudspeaker sound quality evaluation

Alex Karandreas
Department of Acoustics
Aalborg University
Fredrik Bajers vej 7-B5
DK-9220 Aalborg
Revised November 1, 2010

Preface

This thesis is submitted to the Faculty of Engineering and Science at Aalborg University in partial fulfilment of the requirements for the Ph.D. degree. The work has been carried out in the period March 2005 to November 2010 at the Section of Acoustics, Department of Electronic Systems at Aalborg University. During this period, I was employed by the University, and was attached for a period of time with the Sound Quality Research Unit (SQRU) at Aalborg University. The participating companies were Bang & Olufsen, Brüel & Kjær, and Delta Acoustics & Vibration. Further financial support comes from the Ministry for Science, Technology, and Development (VTU), and from the Danish Research Council for Technology and Production (FTP). In March 2005 I was enroled in the Ph.D. programme in Electrical and Electronic Engineering, under the supervision of Associate Prof. Flemming Christensen. I would like to thank all the people who supported my work during the completion of this thesis: my supervisor, Flemming Christensen, for his support and patience; all the staff at the section of Acoustics for always being eager to help and motivate me, all the staff at the Sound-Quality Research Unit for the many fruitful discussions and input on my work; Claus Vestergaard and Peter Dissing for invaluable assistance in setting up the experiments. Finally, I would like to thank my parents, my brother and my dearest friends.

Alex Karandreas, Aalborg, November 2010.

Summary

Currently there is a strong interest in multimodal applications. Communication tools like mobile phones and the internet have steadily switched from audio and text applications to high quality audiovisual media. Current product design of electronics, home appliances and vehicles considers the overall properties of the products and not only their basic functionality. Multimodal applications such as virtual reality systems, virtual environment technologies for the acquisition of motor skills, and feedback/emergency alerts are also becoming available. This trend in technology is dictated by the fact that human perception is by nature multimodal. Indeed, any given product is rarely perceived in isolation, but rather judged within a global context which includes information from all modalities. The processing of multimodal information is not known to be a straightforward linear combination of the subjective impression of each separate modality. In fact there are studies that present evidence where the interaction of these modalities gives results that deviate from a linear combination. This evidence indicates that in order to adequately evaluate multimodal products, any judgement should be based on the relevant multimodal information. Independent unimodal evaluations do not take into account the relative importance of each modality or the way the modalities might interact, which can lead to false conclusions.

Available methodologies for multimodal experiments are novel and not thoroughly validated, unlike the widely accepted techniques for unimodal evaluations. Furthermore, current research in this field is scattered among many disciplines; depending on the scientific discipline and the research goals of each study there are differences in experimental design, stimuli selection and stimuli presentation.

The aim of this thesis is to investigate the relative importance of audio and visual information in subjective evaluations of a product. A multimodal setup was developed in a manner that allowed the subjective audiovisual evaluation of loudspeakers under controlled conditions. A series of music excerpts were reproduced through a loudspeaker while subjects were shown a range of loudspeakers and evaluated the combined audiovisual presentation. The setup and methodology were designed in a manner that would allow the audiovisual evaluation of actual products. Additionally, unimodal audio and visual evaluations were used as baseline tests used for comparison. The experiment provided evidence that unimodal evaluations can be misleading in the relative importance of each modality with respect to the overall quality evaluation. The results show that this was not due to specific interactions between stimuli but rather because the auditory modality dominated over the visual modality.

The same procedure was applied in the investigation of the validity of presenting

substitutes rather than the actual product, in order to assess the necessary level of realism required in multimodal evaluations. In a series of experiments the stimuli presentation consisted either of photographs projected on a 1:1 scale with respect to the actual loudspeakers coupled with audio reproduction through loudspeakers, or small scale-photographs combined with audio reproduction through loudspeakers or small scale-photographs combined with audio reproduction through headphones. All experiments showed similar results and the overall conclusions drawn were comparable to the experiment with actual loudspeakers, thus indicating that the use of substitutes instead of actual products is valid.

Since the experiments revealed the dominance of one modality in the overall evaluation, it was important to rule out potential sources of bias. One potential source of bias was the experimental question, which could have biased attention towards the auditory modality. Therefore, a neutral experimental question was used in identical experiments, where it was concluded that the experimental question had a small but statistically significant effect.

Finally, a different experimental design was investigated in a experiment where each participant evaluated only a subset of the range of stimuli. This was done in order to present subjects with products that had stable characteristics throughout the experiment. In this experiment combinations of loudspeaker models and music excerpts were unique for each subject. The conclusion of this experiment was that the influence of the visual modality was stronger (compared to the previous experiments) while the auditory modality remained the most influential modality.

Resumè (summary in Danish)

For tiden er der en stærk interesse i multimodale applikationer. Kommunikationsværtøjer som mobiltelefoner og internettet har ændret sig fra audio og tekstapplikationer til højkvalitets audiovisuelle medier. Produktdesign af elektronik, hvidevarer og køretøjer tager produktets overordnede egenskaber - ikke kun deres basal funktionalitet i betragtning. Multimodale applikationer som virtual reality systemer, virtuelt-miljøteknologier til læring af motoriske evner, og nødsalarmsystemer, bliver ogsåtilgænglige. Denne tendens omkring teknologen er drevet af den kendsgerning, at den menneskelige opfattelse naturligt er multimodal. En givent produkt er faktisk sjældent opfattet alene, men nærmere vurderet i en sammenhæng, som inkludere informationer fra alle modaliteter. Processeringen af multimodale informationer er ikke kendt for at være en direkte lineær kombination af hver enkelt modalitets subjektive indtryk. Der er faktisk studier som beviser, at interaktionen af disse modaliteter afviger fra en linear kombination. Dette bevis indikerer, at for påtilstrækkelig vis at kunne vurdere multimodale produkter, skal enhver vurdering baseres påden relevante multimodale information. Selvstændig unimodal vurdering tager ikke højde for hver modalitets relative vigtighed eller den måde hvorpåmodaliteterne interagerer. Dette kan føre til falske konklusioner.

Tilgængelige metoder for multimodale eksperimenter er nye og ikke validerede, hvorimod teknikkerne for unimodale vurderinger er bredt accepterede. Endvidere er nuværende forskning i feltet delt over flere fag; afhængig af videnskabelig disciplin og forskningsmål er der forskelle i eksperimentielt design, valg af stimuli samt præsentation af stimili.

Målet med dette studie er at undersøge den relative vigtighed af auditive og visuelle informationer i den subjektivt evaluering af et produkt. Der blev udviklet en multimodal opstilling sådan at subjektive audiovisuelle evalueringer kunne laves under kontrollerede forhold. En serie af musikklip blev spillet over en højttaler mens forsøgspersoner blev forevist en række højttalere og derpåskulle evaluere den kombinerede audiovisuelle præsentation. Opsætningen og metodologien blev designet sådan at faktiske produkter kunne blive audiovisuelt evalueret. Unimodale audio og visuelle evalueringer blev ogsåbrugt som sammenligning. Forsøget viste at unimodale evalueringer kan være vildledende med hensyn til hver modalitets relative vigtighed i forhold til den overordnet evaluering af kvaliteten. Resultaterne viser at dette var ikke pågrund af specifikke interaktioner mellem stimuli men nærmere fordi den auditive modalitet dominerede over den visuelle.

Den samme procedure blev brugt i undersøgelsen af validiteten af at præsentere erstatninger i stedet for det aktuelle produkt, sådan at man kan finde det nødvendig niveau af realisme i multimodale evalueringer. I en række forsøg var stimuli fo-

tografier projekteret i en 1:1 skala (i forhold til de akuelle højttalere) kombineret med lyd gennem højttalere; mindre skala fotografier kombineret med lyd igennem højttalere og mindre skala fotografier kombineret med lyd igennem hovedtelefoner. Alle forsøg viste lignende resultater og de overordnet konklusioner lignede dem fra forsøget med faktiske højttalere. Pådenne måde blev det bevist at erstatninger kunne bruges i stedet for faktiske produkter.

Da forsøgene afslørede at en modalitet dominerede i den overordnede evaluering, var det vigtigt at udelukke mulige biaskilder. En mulig biaskilde var forsøgsspørgsmålet, som muligvis kunne have trukket opmærksomhed mod den auditive modalitet. Et neutralt forsøgsspørgsmål blev derfor brugt i en række identiske eksperimenter hvor der blev konkluderet at forsøgsspørgsmålet havde en lille men statistisk signifikant effekt.

Et anderledes testdesign blev til sidst undersøgt i et forsøg hvor hver forsøgsperson kun evaluerede en undergruppe af alle stimuli, sådan at forsøgspersonerne fik præsenteret produkter med stabile egenskaber hele vejen igennem forsøget. Kombinationerne af højttalermodeller og musikklip var unikke for hver forsøgsperson i dette forsøg. Konklusionerne i dette forsøg var at indflydelsen af den visuelle modalitet var fremhævet mens den auditive modalitet stadig forblev den mest inflydelsesrige.

Contents

1

Introduction

1.1 Area of research

Although cinematography has successfully combined sound and picture to convey feelings, create effects and illusions or to enhance realism, the perceptual effect of combining audio and visual material is still a relatively new research area and there is a lack of a proven experimental methodology: "existing methods for subjective assessment of sound quality are sometimes inadequate for sound systems with accompanying pictures" (ITU-R Rec. BS.1286, 1997).

For audio products e.g. Hi-Fi equipment and other commercial products that emit sound, the use of psychoacoustic evaluation is relatively new (widely used in the last 15 years). It is thus not a surprise that in practise the multimodal evaluation of products is limited. The overall perception of a product may be different when the evaluation is based on a single modality rather than multimodally. In research a subject's attention can change when presented with multimodal or unimodal information, leading to different results. Furthermore, the relation between singular modalities may not be constant, so that under certain conditions, one may dominate the overall quality judgement. Choice of stimuli and the user's expectation of a product might also be influential (Kohlrausch and van de Par, 2005; Woszczyk, Bech and Hansen, 1995). Finally, in the subjective judgement of products, an authentic presentation might be important. The amount, coherence and consistency of information to the different modalities could be a key issue in creating a realistic and natural-feeling presentation (Riva, Davide and Ijsselsteijn, 2003).

1.2 Relevant literature for audiovisual experiments

Models of the human senses intend to describe the main attributes of human perception. Psychometric research has a long tradition for investigations into basic parameters of the human auditory and visual system where relations between stimulus and perceptual response of hearing and vision are investigated and modeled (e.g. loudness, pitch, brightness, contrast, etc). When it comes to higher level qualitative measures, auditory models have been developed to predict subjective speech and audio quality (Thiede et al., 2000; Rix et al., 2002) and models of vision as well as subjective and objective evaluation methods are in use (Takahashi, Hands and Barriac, 2008; Puria, Chen and Luthra, 1995; Winkler, 2005; ITU-T Rec. P.910, 2008; ITU-R Rec. BT.500-12, 2009; ITU-T Rec. J.144, 2004; ITU-R Rec. BT.1683, 2004). These unimodal models however cannot be applied directly to multimodal

investigations since the relationship between modalities is still an object of research.

Audiovisual (AV) research is scattered across many disciplines. Kohlrausch and van de Par (2005) propose a division of this research in certain categories depending on the specific research goals. This division provides a good coverage of all studies related to multimodal experiments with audio and visual stimuli and is also adopted in this thesis. Examples and literature of each category of AV experiments is briefly discussed, while extended information is given on studies focusing on product evaluation that are the most relevant to this thesis. Several ITU recommendations are also presented. These recommendations indicate the effort being made by the scientific community to arrive to solid methodologies for the objective and subjective evaluation of audiovisual products and applications.

1.2.1 AV research for urban environments

This research area examines the relationship between urban noise and other everyday sounds to photographs of landscapes or urban environments, and the associated annoyance perception. Viollon et al. (2002) assessed how listener judgments of a set of urban sound environments were affected by co-occurring photographs of visual urban and rural settings. Subjects rated eight urban sound environments: human sounds (footsteps and voices), bird song, and road-traffic noises when they were associated with five visual settings (four color slides varying in degree of urbanization and a control condition with no slide), along two scales (Unpleasant-Pleasant and Stressful-Relaxing). Results showed that the more urban the visual setting, the more negative the ratings were. However, this result depended on the type of sound. The visual influence was strong for recordings which did not include human sounds, but was absent for all recordings which included human sounds, thus showing that the audiovisual interaction in this case was context dependent. Viollon et al. (2002) offers examples on the applicability of photographs as visual stimuli and produces results that show a non-linear relationship for audio and visual stimuli indicating that the context (semantics) of the stimuli can be important.

1.2.2 AV temporal research

AV temporal asynchrony research has provided the movie and television industries with guidelines on maximum tolerances for the delay between sound and picture (van Eijk et al., 2008). Interestingly, this delay tolerance is not equal for the two modalities but is shown to be larger when the visual stimulus precedes the audio stimulus. Furthermore, research in the field has produced evidence of audiovisual temporal interactions: auditory perception has a pronounced influence on visual temporal rate perception. For example, the temporal rate of the presentation of pure tone stimuli is shown to significantly affect the temporal rate perception of light flashes (Recanzone, 2003). Temporal asynchrony is an issue that was considered during the setup of all experiments presented in the thesis. However, in these experiments the presentation of the visual stimuli was “static”. Thus, the AV synchronization requirement was time-aligned start and stop times for each AV presentation.

1.2.3 AV spatial research

Studies in the field have documented significant interaction effects, where the auditory spatial position is strongly influenced by the visual spatial position, like ventriloquism[1]. Also in cases that do not involve speech perception, when the information from the two modalities is conflicting, the visual modality is the most dominant one (Pick et al., 1969). The effect of visual stimuli on auditory distance perception has shown that visual stimuli can be very influential, even for small distances (Brown et al, 1998). For this thesis it was important the that the audio and visual stimuli would be perceived as a single event, it was thus important that the stimuli would appear to originate from the same spatial location, even when that was not the case. Therefore, for all audiovisual presentations the distance between the spatial origin of the audio and visual stimuli was kept to the practical minimum.

1.2.4 AV speech research

AV speech research refers to studies that investigate the effect of visual information on speech quality assessment or intelligibility. This research area usually deals with the perception of speech while a head and torso video (or a photograph) of a talking person is shown. The McGurk effect[2] (MacDonald and McGurk, 1978) is an essential part of this research and has provided clear evidence for audiovisual interactions. Interesting topics include errors produced by asynchrony and semantic differences between audio and visual stimuli, for example visual information from talkers other than the audible talker. This research area has given evidence of an underlying mechanism or interaction between the modalities and also shown the benefits of a talking image in the presence of noise/low intelligibility situations.

1.2.5 AV and multimodal attention research

Studies like Alais et al. (2006) describe experiments in attention performance in the context of multimodal information focusing on vision and audition. Such research complements other studies in *physiology*, (Brefczynski and DeYoe, 1999) *neurology* (Calvert, Campbell and Brammer, 2000), *channel theories for multimodal information* (Allport, Antonis and Reynolds, 1972) in explaining how the brain deals with multimodal information. Some of these studies suggest that at least for low-level tasks such as discrimination of pitch and contrast, each sensory modality is under separate attentional control, rather than being limited by a unified attentional resource. On the other hand there are studies showing that conflicting multimodal information can deteriorate subjects performance on low level tasks (Spence, Ranson and Driver, 2000), (Taylor, Lindsay and Forbes, 1967), (Massaro and Warner, 1977). Additionally, there are studies that give evidence of "multisensory neurons" in many areas of the brain, which can convey both within- and across-modality information. These neurons respond maximally to interactions of stimuli from more

[1]The skill to speak without opening the mouth while moving a puppet (including its mouth) in order to create the illusion that the puppet is talking.

[2]The McGurk effect is a perceptual phenomenon which demonstrates an interaction between hearing and vision in speech perception. It suggests that speech perception is multimodal, that is, that it involves information from more than one sensory modality. This effect may be experienced when a video of one phoneme's production is dubbed with a sound-recording of a different phoneme being spoken. Often, the perceived phoneme is a third, intermediate phoneme.

than one modality. In this way, multiple sensory cues are not only integrated, but also transformed, producing a reaction larger than that which would be expected from a sum of its parts (Stein and Meredith, 1993). Studies in this field often use artificial stimuli and ask subjects to evaluate pitch, contrast or other specific characteristics of each modality. These studies give evidence for underlying mechanisms however their results should not be generalized.

1.2.6 Multimodal integration research

The evidence from a variety of studies (Spence, 2007), (Hollier and Voelcker, 1997) supports the view that a number of different factors, both structural (spatial, temporal) and cognitive, conjointly contribute to the multimodal integration of auditory and visual information.

According to Hollier and Voelcker (1997) important factors that can influence audiovisual scenarios are timing (sequencing, synchronization), quality balance between modalities, whether the audible or visible error is judged to be important in relation with a specific task or application, high-level cognitive preconceptions associated with the task, attention split, degree of stress introduced by the task (level of difficulty) and the experience of user (novice versus expert). Bearing these in mind, one should be careful when generalizing the conclusions of multimodal studies.

Two interesting theories have been presented: *Semantic congruency:* multimodal integration of audio and visual stimuli is enhanced when AV stimuli are semantically congruent (dog barking sound with picture of a dog) as compared to when they are semantically incongruent or presented unimodally (Molholm et al., 2004), (Laurienti et al., 2004). *The unity assumption:* when the auditory and visual sensory inputs are perceived as being highly consistent (as being related in a way that they appear to go together), observers will be more likely to treat them as referring to a single audiovisual event. Observers will more likely assume that the sensory inputs have a common spatiotemporal origin and more likely bind them together into a single multimodal event, rather than multiple separate unimodal events (Vatakis and Spence, 2008).

For this research the stimuli were semantically congruent and an effort was taken to have common a spatio-temporal origin in order to enhance multimodal integration.

1.2.7 AV product research

AV product studies deal with the audiovisual properties of common household devices including mobile phones, televisions and other multimedia devices, as well as devices that do not reproduce sound, but emit some kind of sound or noise. In a way the AV product category can be thought of as a continuation of sound quality, in the sense that it can be used to predict sound quality or to evaluate sound quality while setting the sound in a specific context (e.g. the soundtrack together with the movie scene or game, the noise together with the vacuum cleaner or car engine), thus providing a more accurate representation to the product at hand. It could be argued that there is a better overall description of the product when sound quality

is complemented by visual quality.

The study by Beerends and de Caluwe (1999) is very significant for this thesis and one of the most important papers in the field. Beerends and de Caluwe propose a series of tests to evaluate the multimodal and unimodal effects of products under test while at the same time instructing the subjects to pay attention to either or both modalities. One of the unique things about this study is the 5 tests AV(AV), AV(A), AV(V), A(A), V(V) where the parenthesis is the modality that the subjects should focus at. Furthermore, from the outcome of the tests Beerends and de Caluwe propose models that define the influence of each of the modalities. The results of this study show that for television commercials the influence of the visual stimuli is greater then that of audio.

Hands (2004) is a study consisting of 2 experiments trying to answer various methodology issues as well as to compare results to the study by Beerends and de Caluwe (1999). The study offers evidence that humans integrate information with a multiplicative rule, implying AV interaction. It also argues that a continuous and categorical rating scale are equally valid and useful. The author acknowledges the fact that the nature of the test material as well as the specific stimuli used in an AV experiment can shift the subjects focus of attention between modalities and strongly influence results and hence any calculated predictive model. The stimuli in the 1^{st} experiment are two videos showing male and female persons talking, while for the 2^{nd} experiment one of the speech videos used in the 1^{st} experiment and another of a bicycle race with audio commentary. In both experiments the original audio and visual stimuli as well as degraded versions are presented. For the 1^{st} experiment the influence of the audio stimuli are slightly greater than that of visual stimuli while in the 2^{nd} experiment the influence of the visual stimuli is dominant. The conclusion is thus that the form of the predictive model is determined by the content of test material under consideration and should only be used in relation to the specific material.

Hollier et al. (1999) presents 3 experiments as well as a comparison among them. The experiments include one that investigates AV temporal asynchrony and two experiments that investigate AV quality perception: a Virtual Reality fly-through of a building with audio commentary and a AV speech experiment. The AV temporal asynchrony experiment validates existing results published by several studies. More interesting are the results of the Virtual Reality experiment that show audio stimuli to dominate the overall quality perception and visual stimuli to have little influence and those of the AV speech experiment showing results where both audio and visual modalities are significant and influential to the overall perception. Interestingly, the design and methodology of the two latter experiments was selected to be exactly the same, thus allowing the researchers to investigate the importance of stimuli in AV experiments. Since the stimuli is the only difference between the two experiments the authors suggest that it is the type of stimuli and the semantic congruency between AV stimuli that causes the difference.

The work by Zielinski, Rumsey and Bech (2003) mainly addresses audio quality, but features an interesting audiovisual section investigating whether a simultaneous video presentation has an effect on the perceived audio quality of a 5.1 multichannel audio system. The stimuli are 12 audiovisual excerpts from movies or concerts, thus actual audiovisual material (not artificial combinations) with the same contextual

content. Results showed that video presence had a negligible although statistically significant effect on the audio quality assessment. The fact that the experimental question was *basic audio quality* (defined as a global attribute describing any and all audio differences), together with the fact that the audio excepts were degraded while the visual excerpts were not, could have influenced results. The paper includes a very thorough section on the related statistics with issues that are commonly encountered in subjective tests but seldom reported. Also interesting is an additional analysis on the audiovisual data under the hypothesis that video presence may only affect the audio quality evaluation of slightly impaired audio excerpts. The suggested statistical techniques were considered during this thesis and a similar investigation on slightly impaired items was carried out (presented in section 4.3 of this thesis).

Bech, Hansen and Woszczyk (1995) investigated the influence of television screen size in the evaluation of a home theater system. The statistical analysis showed that the screen size was a significant factor, and that an increasing screen size resulted in more favorable audiovisual evaluations. Unfortunately, the influence of screen size on overall AV quality has received very little attention - although as the authors point out, practical experience from home cinema and movie theaters shows that up to a point a large format is favourable.

The overview of AV research in Kohlrausch and van de Par (2005) covers in detail spatial and temporal AV interaction, AV distance perception, AV attention, AV speech and includes many of the most important AV product studies. Sensitivity to temporal asynchrony in AV stimuli is thoroughly discussed noting that the naturalness of the stimuli can significantly affect the outcome of the experiment, and the same can happen with the experimental design. For AV reproduction systems the authors summarize saying that research shows a clear mutual influence of the audio and visual modality on the perceived quality: "*The question of which of the two modalities contributes more to the overall AV quality cannot be answered un-equivocally. Depending on the choice of stimuli, i.e. video or still pictures and/or causal relation between audio and video being present or absent, very different amounts and even different signs for the across-modal influence have been found*".

1.2.8 ITU recommendations

In general, ITU recommendations provide a good fundamental source of information for audiovisual experiments, with guidelines for most things relevant to the experiment, including the audio and visual stimuli selection, the experimental setup, the experimental method and assessment techniques (data analysis) and more. Among many ITU recommendations there is some overlap, however each presents some more detailed information concerning each specific topic of interest. Some of the most relevant ITU recommendations are described here with a note on the specific issue that the recommendation deals with:

ITU-T Rec. P.910 (2008), "Audiovisual quality in multimedia services - Subjective video quality assessment methods for multimedia applications.". This document describes test methods and appropriate experimental designs for audiovisual scenarios with video stimuli. The document also describes appropriate viewing conditions and statistical analysis.

ITU-R Rec. BS.1286 (1997), "Methods for the subjective assessment of audio systems with accompanying picture." This document describes suitable setups for viewing and evaluating still pictures when combined with audio, appropriate design methodology and data analysis techniques.

ITU-R Rec. BS.1116-1 (1997), "Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems." This recommendation describes the properties of the listening room and playback requirements, as well as design methodology and analysis issues, including normalization techniques that have been used in this thesis.

ITU-R Rec. BT.500-12 (2009),"Methodology for the subjective assessment of the quality of television pictures." This document describes suitable setups for viewing and evaluating motion and still pictures and the typical measurements to assess the viewing conditions.

ITU-R Rec. BS.775-2 (2006), "Multi-channel Stereophonic Sound System with or without Accompanying Picture." This document recommends appropriate audio setups for subjective evaluation with or without an accompanying picture.

ITU-R Rec. BS.1534-1 (2003), "Method for the subjective assessment of intermediate quality level of coding systems." This recommendation mentions possible methods to degrade an audio signal and suggests appropriate experimental questions.

1.3 State of the art

This chapter's presentation of audiovisual studies indicates the current state of the art in research in this broad field. Relevant literature is limited and scattered across disciplines and journals, while AV studies employ different experimental methods, stimuli and procedures. These studies have produced significant evidence for the usefulness of multimodal investigations. The ITU recommendations reflect current methodologies for audiovisual evaluations. The ITU recommendations are however general approaches to a variety of problems and have certain shortcomings. Interestingly, there are very few studies concerning the audiovisual evaluation of actual products. Even for televisions and home theater systems there exist only a couple of studies and there is still no clear evidence for the relationship of screen size, visual quality and audio quality to the overall audiovisual quality perception.

1.4 Goals of the thesis

The goal of this thesis was to investigate audiovisual interactions by studying the influence of visual appearance on loudspeaker sound quality evaluation. Specifically, this research aimed to answer the following questions:

- in an audiovisual presentation, what effect (if any) does visual quality have on the perception of audio quality and vice versa?

- what is the relative importance of each modality to the overall multimodal evaluation?

- is the audiovisual presentation evaluated to be different than the "baseline" conditions derived from audio-only and visual-only quality evaluations?

- is the relation between modalities linear or non-linear?

The necessary tools for this research were to: (1) establish a suitable multimodal experimental setup (2) verify that listeners could consistently evaluate audio, visual and AV stimuli (3) quantify the evaluations on meaningful scales, and determine the relation of each modality to the overall evaluation (4) test whether the same conclusions are made for experimental setups that differ to some degree, in either the (4a) experimental question, (4b) the audio presentation, (4c) the visual presentation and (4d) the experimental design.

One of the necessary tools of this project was to come up with a valid methodology for the subjective assessment of products with audiovisual characteristics. The methodology used should ideally be valid for a range of products or applications. Therefore, the aim was to select and test methodologies and presentation techniques that would simplify any design while allowing for a reasonable level of information to by extracted by the experiment. However, it was not a goal for this project to design a state of the art audiovisual system or to establish a model for customer preference for the relation between loudspeaker sound quality and loudspeaker cabinet design (a process that would require a large sample of loudspeakers, various music/speech samples and different subject groups controlled for age, sex, nationality, etc.).

1.5 Organization of the thesis

The main part of the thesis consists of five manuscripts, some of which are revised versions of conference papers, and some being intended for publication in peer-reviewed journals. The first manuscript deals with the choice of suitable audio and visual stimuli and the creation of a valid experimental setup. The experiment described in this manuscript investigates the visual influence of actual loudspeakers on subjective audio evaluation. The second, third and fourth manuscripts aim to reveal the importance of stimuli presentation techniques in audiovisual experiments. The fourth manuscript also investigates the effect of the experimental question in audiovisual experiments. In the fifth manuscript a different design methodology is investigated. In the following section, each of these manuscripts is introduced and summarized.

Manuscript A Karandreas, A. & Christensen, F. (2010). "Influence of visual appearance on loudspeaker sound quality evaluation". Submitted to the Journal of the Audio Engineering Society. Parts of this work appeared on: "Influence of visual appearance on loudspeaker sound quality evaluation", 124^{th} *Convention of the Audio Engineering Society, Amsterdam, 2008.*

Manuscript B Karandreas, A. & Christensen, F. (2010). "Subjective audiovisual evaluation with an accompanying large-scale photograph". Submitted to the Journal of the Audio Engineering Society.

Manuscript C Karandreas, A. & Christensen, F. (2010). "Subjective assessment of loudspeaker reproduction with accompanying small-scale photograph - an audiovisual experiment". Submitted to the Journal of the Audio Engineering Society.

Manuscript D Karandreas, A. & Christensen, F. (2010). "The influence of the experimental question in audiovisual experiments". Submitted to the Journal of the Audio Engineering Society.

Manuscript E Karandreas, A. (2010). "Subjective audiovisual assessment of loudspeakers". Submitted to the Journal of Applied Acoustics.

1.6 Interrelations of the manuscripts

Manuscript A : "Influence of visual appearance on loudspeaker sound quality evaluation".

Manuscript A documents the selection of audio and visual stimuli that are also used in subsequent experiments. The experimental setup and procedure is presented along with all decisions that led to the final form of the experiment together with data collected along each step. The main experiment in this manuscript is one where the visual stimuli were actual loudspeakers and the audio stimuli were reproduced by a loudspeaker. The results of the experiment show a dominating influence of audio over visual.

Manuscript B : "Subjective audiovisual evaluation with an accompanying large-scale photograph".

One of the parameters under investigation in this research was different stimuli presentation approaches. Starting with manuscript A the presentation is the actual product, an optimal presentation. In manuscript B onwards the presentation of the audio and visual stimuli gradually deviate from the optimal presentation. Thus, in manuscript B the audio part was maintained while the visual part was a 2D instead of a 3D presentation at a 1:1 scale of the actual loudspeaker. The results presented in manuscript A show little difference between the 2 audio excerpts. Thus in manuscript B it was decided to limit the number of audio stimuli to half. The results were comparable to the data in manuscript A.

Manuscript C : "Subjective assessment of loudspeaker reproduction with accompanying small-scale photograph - an audiovisual experiment".

In manuscript C the audio presentation was the same as in manuscript B, while the visual presentation was a much smaller 2D presentation. This was the least optimal visual presentation in all experiments in the sense that the visual medium was a photograph of the product shown in a scaled-down version. This visual presentation was thought to be a practical way to communicate the information in real life as well as in laboratory experiments. The data was essentially similar to the data from manuscript A and B. Therefore this visual presentation was chosen for subsequent experiments.

Manuscript D : "The influence of the experimental question in audiovisual experiments".

Manuscript D discusses two separate but almost identical experiments. In each the visual part was maintained (small 2D presentation), while the audio was reproduced through headphones instead of loudspeakers. Headphone presentation in this context can be thought as a less realistic approach since the experimental question refers to loudspeakers and the visual stimuli portray loudspeakers. The point of using headphones is that they are more practical in certain applications and could be a potentially useful tool for further experiments. The two experiments only differ in the experimental question. The reason for using different experimental questions was to investigate any possible bias that could focus subjects attention towards one of the modalities. The data from both experiments show that the effect of the experimental question in the context of these experiments is small and that results remain essentially similar to those of the previous experiments.

Manuscript E : "Subjective audiovisual assessment of loudspeakers".

Manuscript E investigates a different design approach as well as a different set of audio stimuli. The goal of this experiment was to limit any possible factors that could potentially bias attention towards the audio part, so that any influence from the visual stimuli would be easier to detect. The resulting data did in fact show a significant influence by the visual stimuli but also verified trends shown in previous experiments, including that the influence of audio in the overall evaluation remained larger than that of visual.

2

Discussion

This section discusses issues that are common across all manuscripts: considerations for stimuli selection in audiovisual experiments and the reasons for selecting the specific rating scale and experimental question. Also, a comment is given on the differences in methodology and stimuli presentation between experiments.

2.1 Stimuli

Stimuli constitute an important parameter in AV experiments, since for the same product or application they have the potential to significantly alter the outcome of the study. Depending on each study's goal, researchers use the appropriate stimuli, but it is important to point out that the same study might lead to different conclusions with a set of stimuli of different characteristics. For AV experiments this means that some established effects are dependent on the type of stimuli and should not be expected to apply universally. For example, in AV speech studies researchers employ phonemes as audio stimuli and examine how they are perceived when they are presented simultaneously with a video of a talking person, with the known results of the McGurk effect. This strong AV interaction is strictly valid for human speech and absent for other stimuli.

The section on AV product studies shows that depending on methodology and stimuli selection the influence of either modalities may be altered.

A number of studies show that video has a much stronger effect than photographs. For example in (Hollier, Rimell, Hands and Voelcker, 1999) the same experimental procedure leads to different conclusions when stimuli with different properties are used (e.g photographs-video, music-speech).

There are few studies that design experiments to investigate the influence of stimuli themselves. In two different studies the authors have shown that the stimuli directly influence the results. In one study (Vatakis and Spence, 2007), audiovisual congruent speech material was shown to have a high temporal asynchrony threshold while incongruent material had a significantly lower threshold. In another study (Vatakis and Spence, 2008) there was no difference between congruent and incongruent presentations of sounds and photographs of musical instruments. Similar outcomes were reported in in (Viollon et al., 2002) where a significant audiovisual effect was found showing that the more urban the visual setting, the more negative the sound ratings, except in cases when the audio included sounds indicating human activity. In those cases the visual setting had no effect over audio perception.

Signal degradations are commonly used in multimodal evaluations as a simple way of examining interactions, that is, having controlled versions of the same stimuli in one modality that can be combined with stimuli of the other modality to give combinations that are related and controlled. For audiovisual experiments it is very common to degrade the audio signal either by adding distortions, filtering or encoding the signal while for visual stimuli it is common to band-pass filter the video signal or to add distortions (Beerends and de Caluwe, 1999; ITU-T Rec. P.930, 1996). For the experiments in this thesis musical excerpts served as audio stimuli and images of loudspeakers as visual stimuli. The audio stimuli were degraded either by adding harmonic distortion or by high-pass filtering the excerpts. The described types of audio degradation were chosen as suitable for the selected stimuli since such audio degradations are realistic for loudspeakers. No degradation of the visual stimuli took place as this was not practical or realistic.

However, according to Hollier and Voelcker (1997), the quality balance between modalities may be important. This means that having degraded stimuli in one modality adds another level of complexity within that modality. It is possible that this can have an effect on the subjects attention.

The stimuli used in this thesis were proven to be well selected when presented in isolation. The pilot tests for the audio and visual stimuli presented in manuscript A showed a clear ranking in both cases, with ratings covering a large part of the rating scale. The audio-only and visual-only tests for each of the experiments verified this. On the other hand, the results of experiments in manuscript A, B, C and the 1st experiment in manuscript D show that the audio modality dominates the audiovisual presentations. This result can be partly attributed to the fact that the audio stimuli were degraded while the visual stimuli were not, and that this extra feature caused subjects to focus attention towards the audio modality. The results of the experiments in manuscript E neither verify or deny this conclusion as the influence of audio to the overall perception is still larger than the influence of visual, but not as dominating.

2.2 Experimental design

2.2.1 Rating method and rating scale

An absolute category rating method was used in all experiments. That meant that each unimodal or multimodal stimulus was presented in isolation and then rated by the subjects. The rating scale used in this study is an adaptation of the conventional 5-point rating scale with labels at midpoints that is recommended in several ITU documents (ITU-T Rec. P.910). It bears features of both the conventional rating scale and a semantic scale as it has labeled anchors, midpoints without labels and is also extended to 9 points. Another significant feature is that the scale is discrete. This scale was chosen because it was straightforward for the subjects to use without having to mind about precision in their answers. Using a similar but continuous scale would have provided with data that could be analyzed with ANOVA without any further considerations, and might have been more normally distributed (since the ratings would not be integers and which could possibly further result in larger spread close to the end points of the scale).For this research it was important to make the presentation and the listening environment as realistic as possible and

thus to interfere as little as possible with the audiovisual environment presented to the subjects. The selected rating scale is a rather non-intrusive way to collect data, contrary to rating scales that would require to select a point on a continuous scale.

2.2.2 Experimental question

In sound quality evaluation, generic terms like "quality" or "basic audio quality" are often used. Quality as a term is however ambiguous (Blauert and Jekosch, 1997; Winkler 1999). On the other hand according to Hollier et al. (1999), "*Information on cross-modal interaction in relation to perceived quality would be invaluable in a system designed to measure the quality of service in multi-modal delivery systems, whereas individual assessment of the individual perceived modality qualities may miss some significant interaction-related problems*". In fact, in multimedia applications (like streaming AV) the term quality is already used for the user settings (users can choose between *low*, *medium* and *high* quality streams in order to select encoding levels that better suit their bandwidth). Finally, the choice of experimental question and scale anchors could be specific instead of generic. For example the subjects could be asked to evaluate the loudspeaker's bass response , the midrange clarity etc., however these terms are specific to loudspeakers and can not be applied to a larger range of products and might be obscure for non-trained subjects. It was thus decided that 2 questions would be used in this research, one of them featuring the term "quality".

It was also interesting to investigate whether the experimental question could be biasing attention towards one modality. To investigate the effect of the experimental question, in 2 experiments (both presented in manuscript D) the question was changed from "How does this loudspeaker sound" which was thought that could bias attention towards the audio part to "Rate the quality of this loudspeaker" which is neutral. The setup, stimuli and methodology was otherwise kept identical. The experimental question was shown to marginally affect results. A comparison between the AV, audio-only and visual-only data of the 2 aforementioned experiments shows that any differences are statistically significant but quite small.

2.2.3 Experimental design method

The experiments in Manuscripts A to D use a full factorial design, therefore all subjects are presented with all possible audiovisual combinations. The advantages of this design is that it is efficient in evaluating the effects and possible interactions of several factors, which is important in multimodal experiments as well as providing a sufficient number of data that with the appropriate statistical analysis can lead to solid conclusions. On the other hand, for product evaluations the choice of a factorial design when degraded stimuli are used means that subjects are presented with products that do not have stable characteristics. Therefore the audiovisual combinations might be perceived as random. A Latin Square design used in manuscript E ensures that each subject is presented with unique audiovisual combinations, so that each product will appear to have stable characteristics. However, that would mean that a large number of subjects would be required in order to collect a sufficient number of data, something that is very impractical considering the training and screening process each subject needs to undergo. Furthermore, such a design rules out the possibility of studying factor interactions which are important for this

research.

2.2.4 Headphone reproduction

The spatial and frequency content between the headphone experiments and the loudspeaker experiments is different. This is an important difference in the audio presentation mode. Changing from loudspeaker to headphone reproduction is somehow comparable to going from 3D to 2D for the visual modality: in the visual modality there is loss of depth and perspective, while in the auditory modality it is the spatial and directional information that is lost. Otherwise, the acoustics of the laboratory are constant across experiments. It should be also noted that for the experiments with loudspeaker reproduction the reproduction was monophonic while for headphone reproduction the same signal was presented to both channels. These reproduction differences are negligible to the changes caused by the degradations in this study. A feature of the set of audio stimuli for these experiments is that the degradations are not spatial and although the perceived spatial position is different there is no dependency on the number of channels.

Manuscript A

Influence of visual appearance on loudspeaker sound quality evaluation

Alex Karandreas[1], Flemming Christensen[1],

[1] *Department of Electronic Systems, Acoustics, Aalborg University, DK-9220 Aalborg, Denmark*

Correspondence should be addressed to Alex Karandreas (aka@es.aau.dk)

ABSTRACT

The overall audiovisual subjective impression of loudspeakers was evaluated in this study. Audio stimuli of varied degradation were coupled with actual loudspeakers of different visual appearance. Additional experiments where the subjects had to evaluate audio-only or visual-only presentations produced a baseline against which the audiovisual evaluation was compared. Results indicate that the influence of audio stimuli dominated the audiovisual evaluation.

1. INTRODUCTION

For audio products i.e. Hi-Fi equipment and other commercial electronic products that emit sound, the use of psychoacoustic evaluation is relatively new (widely used in the last 15 years) and there seems to be only limited application of multimodal evaluations.

Due to interaction effects, perception can change when input from more than a single modality is presented. This concept is investigated in this paper, together with a discussion of the requirements for a thorough study of modality interaction in the laboratory.

Although the perceptual effect of combining auditory and visual material is studied across many disciplines, it is still a relatively new research area and there is a lack of a proven experimental methodology [1].

This is pointed out in ITU-R recommendation BS.1286 [2], which states: "existing methods for subjective assessment of sound quality are sometimes inadequate for sound systems with accompanying pictures".

Models of the human senses intend to describe the main attributes of human perception. Psychometric research has a long tradition for investigations into basic parameters of the human auditory and visual system where relations between stimulus and perceptual response of hearing and vision are investigated and modeled (e.g. loudness, pitch, brightness, contrast, etc). When it comes to higher level qualitative measures auditory models have been developed to predict subjective speech and audio quality [3], [4] and models of vision as well as subjective and objective evaluation methods are in use [5], [6], [7], [8], [9], [10], [11].

Using a bimodal approach, interaction effects can be observed. The relation between singular modalities may not be constant, so that under certain conditions, one may dominate the overall quality judgement. Choice of stimuli and the user's expectation of a product can also be influential [1], [12]. Furthermore, in the subjective judgement of products, an authentic presentation might be important. The amount, coherence and consistency of information to the different modalities could be a key issue in creating a realistic and natural-feeling environment [13].

This paper describes an experiment that uses loudspeakers as the source of both auditory and visual stimuli. The aim is to create a very natural presentation and also identify problems that arise when using actual products as stimuli. The first part of this paper presents two pilot experiments used to select the audio and visual stimuli. Then the setup of the audiovisual experiment is described and finally experimental results are presented and discussed.

2. METHOD

In order to enable a bimodal study of audiovisual interaction, both audio and visual stimuli should be

applied in a way that the test subjects are exposed to a range of different levels of the two modalities studied. In order to select appropriate stimuli two unimodal pilot experiments were conducted:

1. Visual: The subjects judged presumed sound quality of different loudspeakers based on photographs.
2. Audio: The subjects judged audio excerpts that feature degrading artifacts.

Based on the pilot experiments a combined experiment was set up, in which the real loudspeakers selected from pilot experiment 1 were presented in combination with audio excerpts selected from pilot experiment 2. The resulting audiovisual experiment investigated whether a change of quality in one modality affected the subjects perception of quality in the other modality.

2.1. Selection of visual stimuli and visual-only pilot experiment

A well defined method to create visual stimuli spanning a certain range on a quality scale, is to process original (non-degraded) photographs and create degraded versions that have a marked difference which can be also objectively quantified [9], [14]. However, when dealing with physical objects the situation is different since the physical appearance can not be altered. Instead, a range of products within the same family can be chosen and a selection can be made on the condition of a marked subjective difference between the individual products.

This section describes a pilot experiment, necessary in selecting a number of loudspeakers to be used as visual stimuli in the main experiments . The range of the loudspeakers covered from high-quality models, to standard 2-way systems and down to low quality personal computer loudspeakers as seen in figure 1.

For the pilot experiment and for practical reasons (large number of loudspeakers), photographs of loudspeakers were used. Photographs of 12 loudspeakers were taken in identical background and light conditions. Care was taken to preserve the aspect ratio of the loudspeakers when presenting these photographs to the subjects. Bookshelf loudspeakers were shown on the same loudspeaker stand by

Fig. 1: An ensemble of all the loudspeakers used in the pilot experiment, showing the range of loudspeakers used. Aspect ratio is not maintained in this figure. The range of loudspeaker sizes varied from 12.5 x 9 cm to 184 x 18.5 cm.

means of digital photograph manipulation. To aid subjects in acquiring a correct impression of the actual size of each loudspeaker, the loudspeakers were photographed next to a piece of furniture, as shown in figure 2.

A group of 12 university students, all of them naive[1] subjects (6 male, 6 female, mean age 23.4 ± 1.24 years) participated in the visual pilot experiment. The presentations were randomized and counterbalanced. An absolute category rating method was used. Each subject gave 2 ratings per stimulus.

In the laboratory, the participants were presented with a series of the aforementioned photographs on a 19 inch computer monitor. During the pilot experiment the size of the pictures on the 19 inch monitor was constant with dimensions: 17 x 12 cm. The viewing distance from the screen was fixed to 1m (that is a viewing distance of 3 times the screen height, recommended in ITU recommendation BS.1286 [2]). An example presentation is shown in figure 2. After each picture was presented, participants were instructed to rate the presumed audio quality of the loudspeaker and urged to imagine the sound the loudspeaker produces. The question was "Tell us how you think this loudspeaker would SOUND". The intention of this question was for subjects to consider the audio properties of the illustrated loudspeakers. Therefore this pilot experiment also investigated whether subjects can draw conclusions on the presumed sound quality of the loudspeakers based only on visual information (in other words whether naive subjects have some preconceptions of what high and low quality loudspeakers look like). No audio was presented in this experiment.

Ratings were given on a discrete rating scale as shown in figure 2. The rating scale was developed from recommendations in [8] and [15], and a 9-point scale was preferred over a 5-point scale for higher discriminative power. The end points were defined by the phrases "low quality" and "high quality". Quality is a descriptor that can be used for audio, visual and bimodal perception studies and is recommended by ITU standards [8] and [15]. The task was displayed above the rating scale as shown in figure 2.

[1] In the context of this paper, a naive subject is one that has no prior experience from viewing or listening experiments and has limited knowledge of loudspeakers (technical and commercial).

Fig. 2: Example of a trial during the visual pilot experiment. The photograph of the loudspeaker includes a piece of furniture. The piece of furniture helps subjects to recognize the size differences of the loudspeakers. The question, rating scale and anchors can be seen.

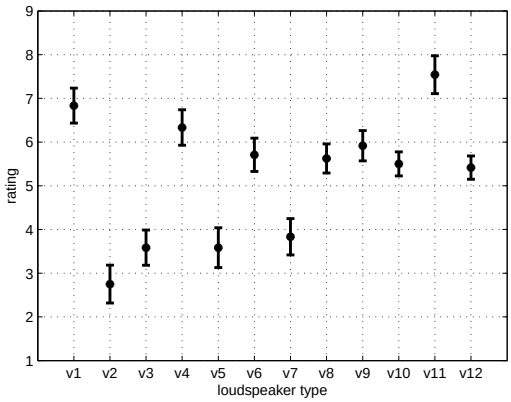


Fig. 3: Ratings of loudspeakers based only on visual appearance. Average ratings ± 1 standard error are shown. The order of loudspeakers is the same as in figure 1 from top left to bottom right. A rating equal to 1 corresponds to low quality and 9 to high quality.

The obtained responses were analyzed with a two-way analysis of variance (ANOVA) for factors *loudspeaker picture* and *subject*. The average ratings are presented in figure 3. The statistical analysis showed a main effect for the *loudspeaker picture* and *subjects* factors ($p < 0.001$ in both cases). Subject differences are to be expected since no reference is used in the experiment, and the subjects receive minimal training.

As seen in figure 3, the standard error bars for loudspeakers *v2*, *v4*, *v7*, *v10* and *v11* do not overlap (this was the selection criterion used). Loudspeaker *v12* has a similar rating to loudspeaker *v10*. The design of the two loudspeakers is also similar (2-way units in a rectangular cabinet), however loudspeaker *v12* has a clearly visible bass reflex hole. Loudspeaker *v10* was preferred because it was readily available. These loudspeakers were chosen to be used as the different levels of visual stimuli in the main experiments. The selected loudspeakers were:

1. Satellite (of a surround system) 1-way unit in grey plastic cabinet. Dimensions: 12.5 x 9 cm. Diaphragm not visible.
2. Large bookshelf 3-way loudspeaker with 4:3 cabinet proportions. Dimensions: 29 x 41 cm. Diaphragm not visible.
3. Large bookshelf 2-way unit with a rectangular wooden cabinet. Dimensions: 35 x 23 cm. Diaphragm visible.
4. Floor standing 4-way loudspeaker. Dimensions: 184 x 18.5 cm. Diaphragm not visible.
5. Small bookshelf 1-way unit in black plastic cabinet with a tilted upper section. Dimensions: 20.5 x 13 cm. Diaphragm not visible.

2.2. Selection of audio stimuli and audio-only pilot experiment

This section describes the selection of audio stimuli to be used in the main experiments, as well as a pilot experiment that was used to verify the appropriateness of the stimuli. The audio stimuli were selected on the principle that they should have a marked difference that can be objectively accounted for and be naturally occurring in loudspeakers systems. One choice was to degrade the original audio by high-pass filtering, as to simulate lack of bass response in small diaphragm loudspeakers. The other was adding levels of harmonic distortion to simulate gross distortions that might occur when loudspeakers are driven beyond their linear operating range. Both methods of degradation are typical in subjective audio evaluations [16]. The degradations were implemented by means of digital signal processing.

Combining non-degraded audio stimuli with some of the loudspeakers would seem unnatural, since very small loudspeakers would be coupled with audio stimuli that featured substantial energy at the low frequencies. Therefore, to ensure that the combination of all audio stimuli and loudspeakers was not unnatural, the original excerpts were high-pass filtered at 110 Hz. All other degradations were made using these already filtered stimuli.

For the high-pass filtering case, Direct form II IIR Butterworth filters were used. The cut-off frequencies were 110, 220 and 440 Hz with a filter order of 15 and a 90dB/octave slope. These stimuli are referred to as: *HP1*, *HP2* and *HP3* respectively.

In order to degrade the excerpts by means of harmonic distortion, a polynomial function that produces 2^{nd}, 3^{rd}, 4^{th} and 5^{th} order harmonics was applied to stimulus HP1:

$$y = x + amp * (-6x^2 - 16x^3 + 8x^4 + 16x^5)$$

where x represents HP1, y the resulting distorted signal and *amp* is a gain that changes the level of the harmonic distortion.

In order to produce stimuli with noticeably different levels of distortion, 4 versions of the HP1 were produced with *amp* set to 0.3162, 0.5623, 1 or 1.1220 respectively for each version. These stimuli are referred to as D-1, D-2, D-3 and D-4 respectively. Care was taken to ensure that no clipping occurred to the resulting signals. The stimuli were not equalized with respect to loudness.

In the laboratory a loudspeaker (Genelec 1031-A) was placed on the listening axis at a distance of 6.7m from the subject. The geometrical center of that loudspeaker was fixed to a height of 1.2m from the floor. The loudspeaker was calibrated to produce a 75dB SPL at the listening position when reproducing 1/3 octave band-limited pink noise with center frequency either at 400 or 1000 Hz (-6dBFS at 44100 Hz, 16 bit). An acoustically opaque curtain hid the loudspeaker from the subjects.

The experimental question was "Please rate the audio quality" with anchors "bad" and "excellent". The rating scale was the same as for the visual-only pilot experiment.

A group of 4 naive subjects (university students, 2 male and 2 female, mean age 23.25 ± 1.5 years) participated in this study and each evaluated all stimuli giving 2 ratings per stimuli. The presentations were randomized and counter-balanced. Each subject gave 2 ratings per stimulus.

The 3 music excerpts used in the pilot experiment were:

1. A rock/country recording [17] featuring an ensemble with drums, bass, guitar and male vocals. Timing [min.]: 0:00 - 0:09.

2. A reggae recording with a strong bass line [18]. The song selection features drums, bass, guitar, keyboards but no vocals. This track is quite sensitive to distortions and features a lot of low frequency content. Timing [min.]: 0:15 - 0:26.

3. A jazz/rock recording featuring a classical guitar with clean mid range sound and a accompanying bassline [19]. Timing [min.]: 0:08 - 0:19.

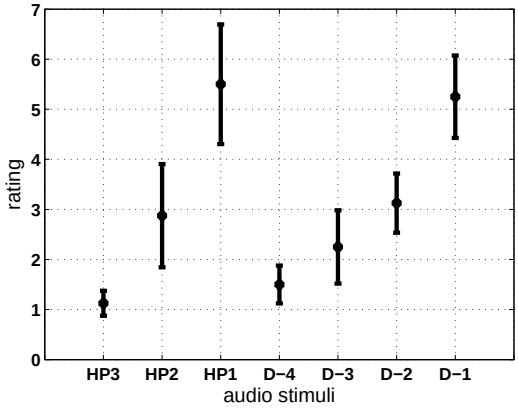


Fig. 4: Results (means ± 1 standard error) of the audio-only pilot experiment for excerpt 1. Stimuli labeled as *HP* are high-pass filtered while stimuli labeled *D* are harmonically distorted. As seen in the figure, the two types of degradation combined cover effectively a wide range of the scale.

The musical excerpts were selected from commercial and reference recordings and transferred (ripped) to a computer (44.1kHz, 16bit). The excerpts had a duration between 9 and 11 seconds and included a complete musical phrase.

The obtained responses were analyzed with a 3-way ANOVA (factors were *excerpt*, *degradation level* and *subjects*). ANOVA results show all factors to be statistically significant ($p < 0.001$ for all cases). Data for one excerpt (typical for all cases) is shown in figure 4.

The pilot experiment revealed a problem with excerpt 3 (the jazz/rock excerpt). When harmonically distorted, the guitar in this track would be perceived very different but not necessarily degraded. Also, because only 2 instruments are featured in excerpt 3, high-pass filtering at 220 and 440 Hz made it sound very "thin" or "hollow" and decreased its sound level

drastically. It was thus decided to exclude this excerpt from the main experiment. Additionally it was decided that the levels of the harmonically distorted stimuli would be reduced from 4 to 3 for the main experiment since the difference between stimuli D-2, D-3 and D-4 was not sufficiently large. The 2^{nd} most harmonically distorted stimulus (D-3) was excluded.

It was also decided that the levels of the distortion components were low. To ensure that clearly audible differences existed between the degraded stimuli, the level of the harmonic distortion was raised. For the rock/country excerpt (excerpt 1) *amp* was set to 0.5623, 1 or 1.2589 respectively for each version. For the reggae excerpt (excerpt 2) *amp* was set to 0.5623, 1.1220, or 1.7783 respectively for each version. Informal listening verified that these levels were appropriate.

2.3. Audiovisual experiment - Presentation of actual loudspeakers as visual stimuli

2.3.1. Setup

In the audiovisual experiment subjects should be exposed to combinations of the audio and visual stimuli described in the pilot experiments (sections 2.1 and 2.2). This imposes some challenges: to present only one visual stimuli at a time, since the loudspeakers are physical objects in the room, and to present the audio stimuli without influence from the reproduction characteristics of the loudspeakers used. The first challenge was handled by making the laboratory room completely dark, and having each set of loudspeakers illuminated independently by means of very narrow spotlights. The second challenge was overcome by using one single hidden loudspeaker to reproduce all audio stimuli.

The experiment was conducted in a laboratory room conforming with the ITU-R BS.1116 recommendation for listening rooms suitable for evaluations of multichannel audio systems [20]. Five pairs of loudspeakers were used, as described in section 2.1. They were positioned in stereo pairs at positions labeled *a, b, c, d* and *e* in the following manner: *a-b-c-d-e-e-d-c-b-a*. Care was taken to align the loudspeakers so that the geometric center of each loudspeaker was at an equal height from the floor (1.2m). Loudspeaker stands were hidden under pieces of cloth that reached the floor. The setup of the loudspeakers can be seen in figures 5, 6 and 7.

However, these loudspeakers do not reproduce sound during the actual test. They only reproduce sound during a familiarization session prior to the actual test. The audio stimuli presentation during the experiments was done by means of a single loudspeaker (Genelec 1031-A) centered behind all other loudspeakers. This loudspeaker was mounted in a fake wall, and covered by a fabric surface and was thus invisible to the subject (see figure 5). An additional acoustically transparent black curtain (placed flat against the wall) was used to minimize light reflecting from the wall. The direct path from loudspeaker to listening position was acoustically unobstructed. The loudspeaker was calibrated to produce a 75dB SPL at the listening position when reproducing 1/3 octave band-limited pink noise with center frequency either at 400 or 1000 Hz (-6dBFS at 44100 Hz, 16 bit). The loudspeakers used as visual stimuli were calibrated with 1/3 octave band-limited pink noise at 500Hz, 1kHz and 2kHz (-6dBFS at 44100 Hz, 16 bit) and adjustments were made to the individual gain of each pair of loudspeakers so that the SPL was 50dB $\pm$ 2dB at the listening position for each loudspeaker.

The experiment was controlled by a PC running custom-made software (programmed in Labview 6.1), situated in an adjacent control room. This included the order of presentation of the stimuli (randomization), the audio stimuli generation, control of the spot lights and the data collection. The audio stimuli were generated by an internal sound card (RME Digi 9636, 24-bit, 96 kHz) were converted to analog (Tracer Technologies Big Daddi, 24 bit D/A converter) and reproduced by an active loudspeaker (Genelec 1031-A, Free field frequency response: 47-22000 Hz ($\pm$ 3dB), crossover frequency: 2.2 kHz, short term RMS @ 0.5m $>$ 107 dB SPL).

In order to have a constant frequency and directivity response for all audio stimuli, with a common on-axis path to the listener and keep the influence of room reflections as constant as possible, a monophonic playback was chosen since it presents a simple and valid approach.

During each audiovisual presentation the subjects should only by presented with the intended stimuli. For this reason, the listening room was kept

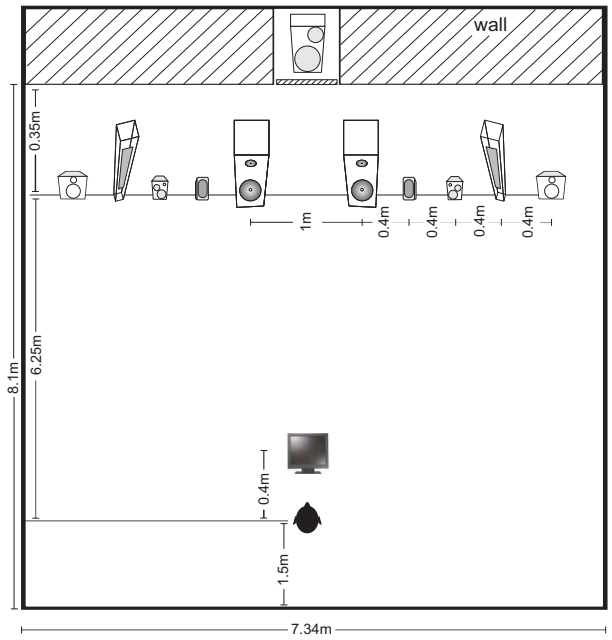


Fig. 5: Top view of the setup in the listening room.

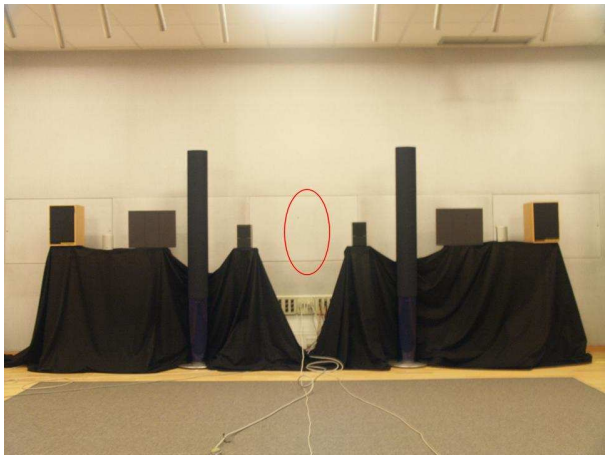


Fig. 6: Photograph of early stages of the setup (room fully illuminated), where the fake wall can be seen and the position of the hidden loudspeaker is marked with an ellipse.

Fig. 7: Photograph of the setup (room fully illuminated). The fake wall is covered with a black cloth. The spot lights that illuminate the loudspeakers during the experiment are also visible. In the foreground, the chair for the subject and the touch screen are shown. The loudspeaker cable connections were left visible for the sake of plausibility.

completely dark. This means that under these conditions subjects could not see the loudspeakers at all. A series of spot lights placed at the ceiling of the room and at a close distance to the loudspeakers were programmatically controlled. Each of them faced a single loudspeaker and could be individually controlled to turn on and off to illuminate just the selected loudspeaker. For all audiovisual presentations the spot lights were synchronized with the onset and offset of the sound. The only other source of light in the room was a touch screen (ELO Touchsystems ETL12IC, 12 inch screen) that displayed the question and rating scale and was the interface by which subjects gave their ratings. The touch screen would switch to black during the audiovisual presentations.

The light conditions in the laboratory room were measured (Gossen MAVOLUX 5032C, Class C acc. DIN 5032-7, Min. Sensitivity: $0.1lx$) . All measurements (table 1) were made at the geometrical center of each loudspeaker and touch screen with all room lighting, spot lights and screen turned off (room completely dark).

2.3.2. Considerations for the realism of the au-

Table 1: Light and viewing conditions in the laboratory room.

Measurement	**luminance** (lx)
Spotlights turned on for loudspeaker pair *a* and measuring at:	
loudspeaker pair *a*	66,3
loudspeaker pair *b*	4,8
loudspeaker pair *c*	0,1
loudspeaker pair *d*	0
loudspeaker pair *e*	0
Spotlights turned on for loudspeaker pair *b* and measuring at:	
loudspeaker pair *a*	1,1
loudspeaker pair *b*	78,2
loudspeaker pair *c*	2,3
loudspeaker pair *d*	0,1
loudspeaker pair *e*	0
At the touch screen with all spotlights turned on	0
Background room illumination	0
Room size	$8.1 * 7.34 * 2.86$ m
Viewing distance to the loudspeakers	6.25 m
Viewing distance to the touch screen	0.4 m

diovisual presentation

In this experiment it was important that participants would accept the notion that the shown loudspeakers did indeed reproduce the audio stimuli. If they came to the conclusion that there existed a separate acoustic source, then it is uncertain whether they would associate the shown loudspeaker with the reproduced acoustic stimuli. It is possible that they would segregate the two modalities and base their judgement on only one of them. In consequence, it was important to support the illusion that the shown loudspeakers were the ones reproducing the audio stimuli. The simultaneous onset and offset of the audio and visual stimuli enhances the perception of a unique audiovisual event [21]. Furthermore, specific instructions for the subjects and a special familiarization session (presented in section 2.3.4) were included before the experiment.

Participants might have expected that each loudspeaker should have unique sound fidelity characteristics. However, each loudspeaker was associated with audio stimuli with varying levels of degradation. This led to a situation where a certain loudspeaker would perform differently within the same experiment. This called for a plausible explanation to the participants. The participants were instructed that: "*Both loudspeakers (in each loudspeaker pair) will be playing the same sound simultaneously and you will get the feeling that the sound is always positioned in the center, in front of you. You might notice that the same loudspeaker produces sounds that vary in quality. This is because of different sound processing methods*".

2.3.3. Subject pre-selection

21 university students (11 male and 10 female) participated in this study (mean age 23.1 ± 2.4 years). An audiometric screening was made, and none had a hearing loss greater than 15dB HL in any ear at any octave band frequency between 125 Hz and 8 kHz. The participants were required to have normal or corrected to normal vision and normal color vision. The participants vision was inspected prior to the experiments using standard vision charts. Concerning acuity, no error on the 20/30 line of the standard eye chart should be made. Concerning color, no more than 2 plates should be missed out of 12 on an Ishihara test [8], [9], [15], [22]. No subjects were

excluded due to failed visual screenings.

As a further supplement, prior to the experiment, data regarding the listening habits and prior experience of the participants were collected by means of a questionnaire to ensure that the subjects were naive listeners.

2.3.4. Familiarization

The subjects were not in contact with the setup until the first familiarization. The aim of this familiarization was to present the loudspeakers and to convince the subjects that all presented loudspeakers could in fact reproduce audio stimuli. Therefore, only for this session, the visible loudspeakers reproduced audio stimuli. A pure tone sequence (1kHz, 2kHz and 500Hz) was monophonically reproduced successively by each loudspeaker. While a loudspeaker reproduced the pure tone signal, the respective spot light was turned on.

The subjects did not have any other task than to observe these loudspeakers while they were illuminated and reproduced audio stimuli.

The second familiarization, introduced the range of degradations, presenting the least and most degraded excerpts. In order to present subjects with the range of audio degradations, only during this familiarization the audio stimuli were labeled: the least degraded stimulus was termed "excellent" and the most degraded stimuli (of both degradation methods) were termed "bad". Subjects were instructed that these stimuli were just some of the audio stimuli they would hear during the experiment. The room was kept completely dark.

A third familiarization emulated the stimuli presentation and evaluation procedure as it would take place during the actual experiment. Four audiovisual presentations were given, during which the least degraded stimulus of each excerpt, and the two most degraded stimuli (high-pass filtering for one excerpt, harmonic distortion for the other) were coupled with one loudspeaker. The presented loudspeaker was also counter-balanced among subjects. Subjects were urged to pay attention to the question they would have to answer, and to familiarize themselves with the user interface.

2.3.5. Experimental design

The experiment was divided in 3 parts: an audiovisual (AV), an audio-only (A-only) and a visual-only (V-only) experiment. For the AV experiment, all visual stimuli were combined with all audio stimuli in a full factorial design. The 12 audio stimuli (2 excerpts, each with 3 high-pass filtered and 3 harmonically distorted versions) were paired with the 5 visual stimuli, generating a total of 60 audiovisual combinations. For the A-only experiment there were ratings of the 12 audio stimuli alone, and for the V-only part ratings of the 5 visual stimuli alone.

The AV experiment was always presented first, followed by either the A-only or V-only experiment. The order of the unimodal experiments was counterbalanced across subjects. Prior to the AV experiment the subjects went through the screening for hearing and vision and were familiarized with the stimuli and procedure as already discussed.

In order to reduce positional bias, the physical position of the loudspeakers in the room should be counter-balanced. Ideally this should be done according to a completely counter-balanced design, but for practical reasons (since the setup involves several heavy loudspeakers that would take a long time to move) a setup with only two different settings was chosen. For half of the subjects the loudspeakers were positioned as: `a-b-c-d-e-e-d-c-b-a` and for the other half as: `e-d-a-c-b-b-c-a-d-e`.

For the AV experiment, each subject gave 2 ratings per audiovisual combination. The presentation of the stimuli was randomized and counter-balanced. A rating scale with anchors following the ITU recommendations [8], [15] and [9] was used as described in the pilot experiments. The aim of the AV experiment was to determine the degree to which visual bias influences audio perception. It was crucial that the subjects perceive the audiovisual presentation as an entity, so that the audio stimuli would not be perceived as an isolated phenomenon but an integral part of the loudspeaker presentation. Accordingly, the experimental question was: "How does this loudspeaker sound?".

The same experimental question was used the A-only ratings. During that period the loudspeakers were completely invisible to the subjects. Each sub-

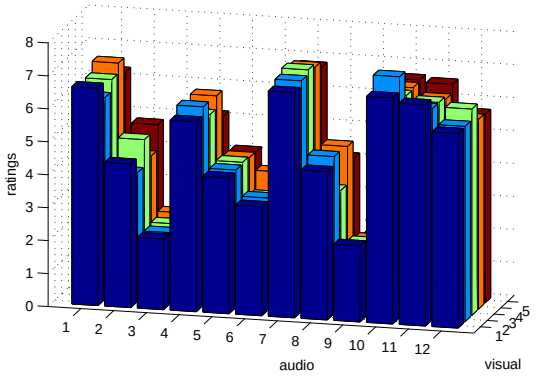


Fig. 8: Average ratings across subjects for all combinations of audio and visual stimuli, AV experiment data. A rating equal to 1 corresponds to low quality and 9 to high quality.

ject gave 2 ratings per stimuli and the presentation order was randomized and counter-balanced.

For the ratings of visual stimuli in isolation, the question was "How would this loudspeaker sound?". Each loudspeaker pair was illuminated for 5 seconds. Each subject gave 2 ratings per stimuli and the presentation order was randomized and counter-balanced.

3. RESULTS

3.1. Audiovisual experiment results

The average results across subjects for all audiovisual combinations are shown in figure 8. Audio levels 1 to 3 and 7 to 9 are high-pass filtered versions of excerpt 1 and 2 respectively, while audio levels 4 to 6 and 10 to 12 are harmonically distorted versions of excerpt 1 and 2 respectively. Large differences are seen on the ratings along the audio axis whereas the ratings along the visual axis exhibit small differences. For the high-pass filtered audio stimuli, the results look similar between the two music excerpts whereas a difference is seen for the distorted audio stimuli, where for excerpt 1 (audio stimuli 4-6) there is a more pronounced decrease in rating for different degrees of distortion while for excerpt 2 (audio stimuli 10-12) the ratings are high and the effect of

Analysis of Variance

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
S	1255.4	20	62.772	36.59	0
A	7065.3	11	642.303	374.42	0
V	10.2	4	2.557	1.49	0.2028
S*A	2714.9	220	12.34	7.19	0
S*V	185.9	80	2.324	1.35	0.0231
A*V	138.2	44	3.14	1.83	0.0009
S*A*V	1637.7	880	1.861	1.08	0.0939
Error	2161.5	1260	1.715		
Total	15169.1	2519			

Constrained (Type III) sums of squares.

Fig. 9: The ANOVA table for the AV experiment including 2 and 3-way interactions. S, A and V refer to factors *subjects*, *audio* and *visual* respectively.

distortion causes a clear but small difference between ratings.

The data were analyzed by means of an ANOVA, shown in figure 9. Factors *subjects* (S), *audio* (A) and all 2-way interactions $S*A$, $S*V$ and $A*V$ are significant. The ANOVA analysis shows that factor *visual* (V) and the 3-way interaction are not statistically significant. The results for the interaction terms $A*V$ and $S*V$ show that although main factor V is not statistically significant it is important to the statistical model.

The ratio of the Sum of Squares of each term to the total Sum of Squares can be expressed as a percentage contribution of each term to the ANOVA model (for more information see [23]). Thus, for this experiment, factor A accounts for 46% of the variability of the experiment, while factor V accounts for 0,07% and the interaction term $A*V$ accounts for 1%.

The 2-way interaction $A*V$ shown in figures 8 and 10 indicates a difference in the evaluation of the harmonically distorted stimuli for the 2 excepts. Indeed, in a 4-way analysis where factor *audio* was split into 2 factors: *music excerpt* (2 levels) and *degradation* (6 levels) the interaction *music excerpt* $*$ *degradation* was statistically significant showing that the effect of degradation was different for the 2 music excerpts, in agreement with figure 8 and 10. The 4-way analysis also showed that the 3-way interaction *subject*degradation*visual* was statistically significant while *degradation*visual* was not statistically significant. That indicates differences in opinion between subjects about certain *degradation*visual* combinations.

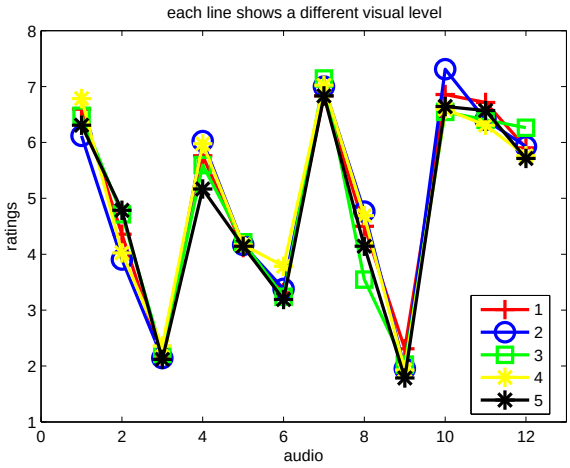


Fig. 10: Average ratings across subjects for the 2-way interaction between factors *audio* and *visual* in the AV experiment. The legend shows the correspondence to the visual stimuli.

The *S*A* and *S*V* interactions are shown in figures 11 and 12. These figures give insight to these significant interactions by grouping subjects with similar response patterns. The criteria used for splitting the subject's responses into subgroups are the general pattern of responses and the overall rating level of each subject.

The results for the *S*A* interaction (figure 11) show a general trend that is followed by almost all subjects, but with some individual variation especially in the cases of subject 6 (*s6*), *s21* (middle left plot) and *s2*, *s11* and *s13* (bottom plot).

Figure 12 shows results for the *S*V* interaction. Most subjects rate the levels of factor *visual* very closely (all within 1 rating point, in some cases 0.5 point) but there are exceptions, *s5*, *s18* and *s19* rate the levels with differences of up to 1.5 rating points. Overall the across-subject differences are greater than the level differences within factor *V*.

Statistical analysis showed that the data from the audiovisual experiment were not normally distributed. A normal probability plot showed that the residuals were not normally distributed but rather had a characteristic *S* shape. Attempts to normalize the data had minimal effect and did not change the shape of the distribution. Non-parametric analysis

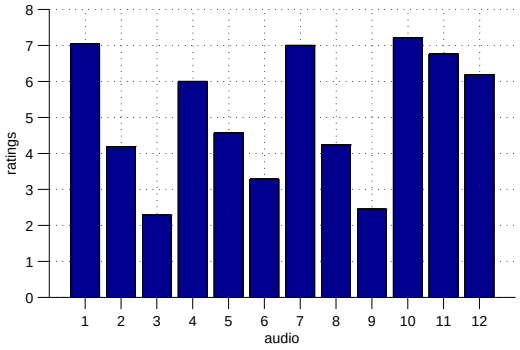


Fig. 13: Average ratings across subjects for the A-only experiment.

Analysis of Variance

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
S	325.63	20	16.282	12.36	0
A	1541.63	11	140.149	106.38	0
S*A	743.37	220	3.379	2.56	0
Error	332	252	1.317		
Total	2942.63	503			

Constrained (Type III) sums of squares.

Fig. 14: The ANOVA table for the A-only experiment.

on the raw data showed results to be very similar to those of the ANOVA (the same terms are statistically significant).

3.2. Audio experiment results

The average results across subjects for all audio-only presentations are shown in figure 13. The figure shows that the are large differences between the ratings of the levels of excerpt 1 (stimuli 1 to 6), while for excerpt 2 the ratings of the stimuli with harmonic distortion show smaller differences. Interestingly, *a10* is rated higher that the least degraded stimulus *a7* while *a11* and *a12* are also rated high. The ANOVA table is shown in figure 14. Factors *subjects* and *audio* as well as the 2-way interaction are all strongly significant.

3.3. Visual experiment results

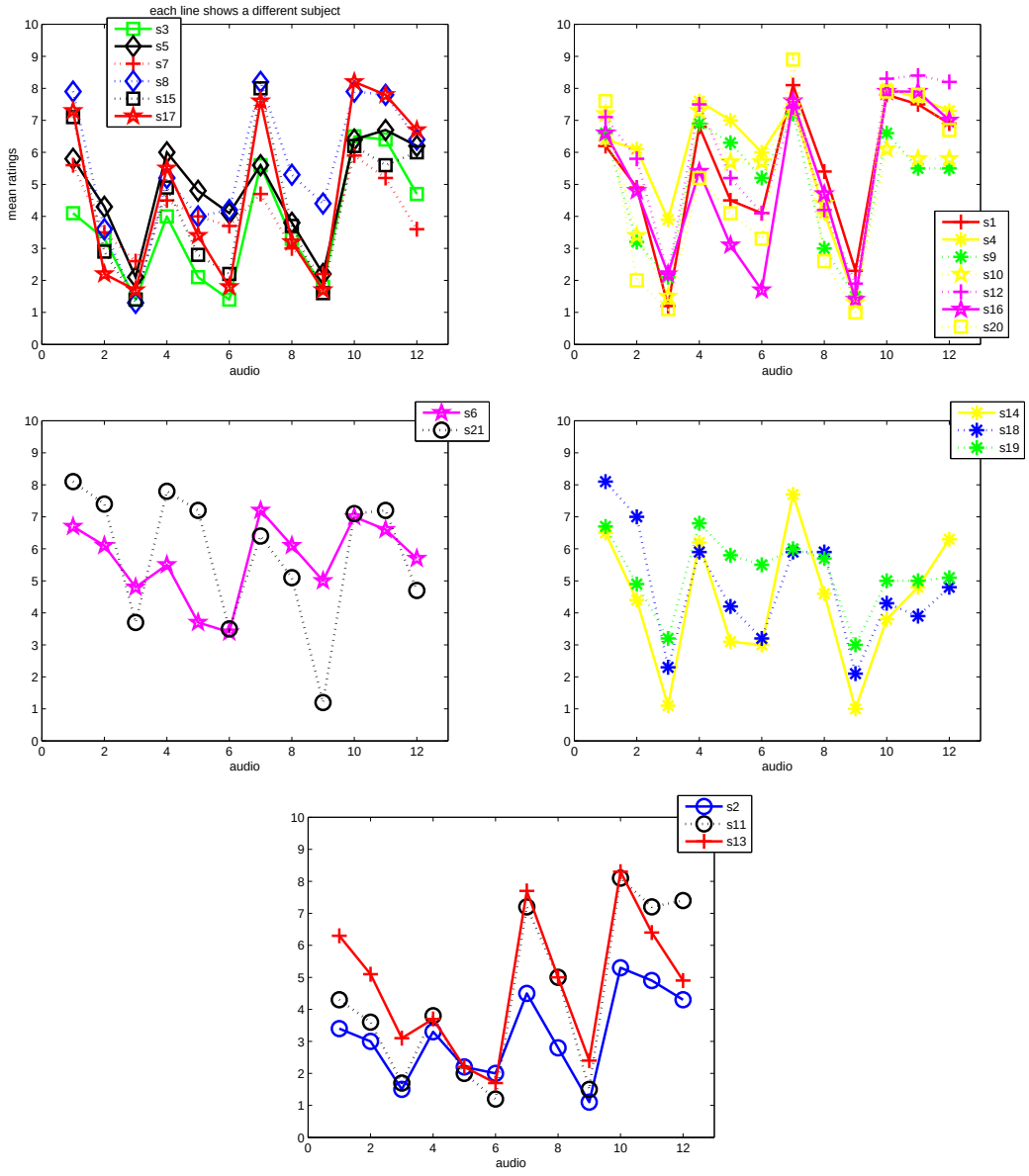


Fig. 11: Plots of means for the *S*A* interaction (AV data) where the individual response of each subject is shown. Subject responses that are similar are grouped together in subplots. The `top left` plot features 6 responses that are similar to the results of the A-only experiment (see figure 13). The `top right` plot features responses that are similar to the `top left` but there is a larger usage of the scale and the harmonically distorted stimuli are elevated showing that this group of subjects feels that harmonic distortion doesn't really impair the quality. The `middle left` plot shows 2 responses that don't fit with the other subgroups. The response for *s21* shows that this subject rates only the most degraded stimuli as low in quality, while the responses for *s6* are condensed to the middle of the scale. The `middle right` plot is similar to the top left but the last 3 responses (harmonically distorted stimuli, excerpt 2) show progressively improved ratings for progressively more degraded stimuli. The `bottom` plot is quite different than the rest with excerpt 1 (first 6 responses) being rated very low.

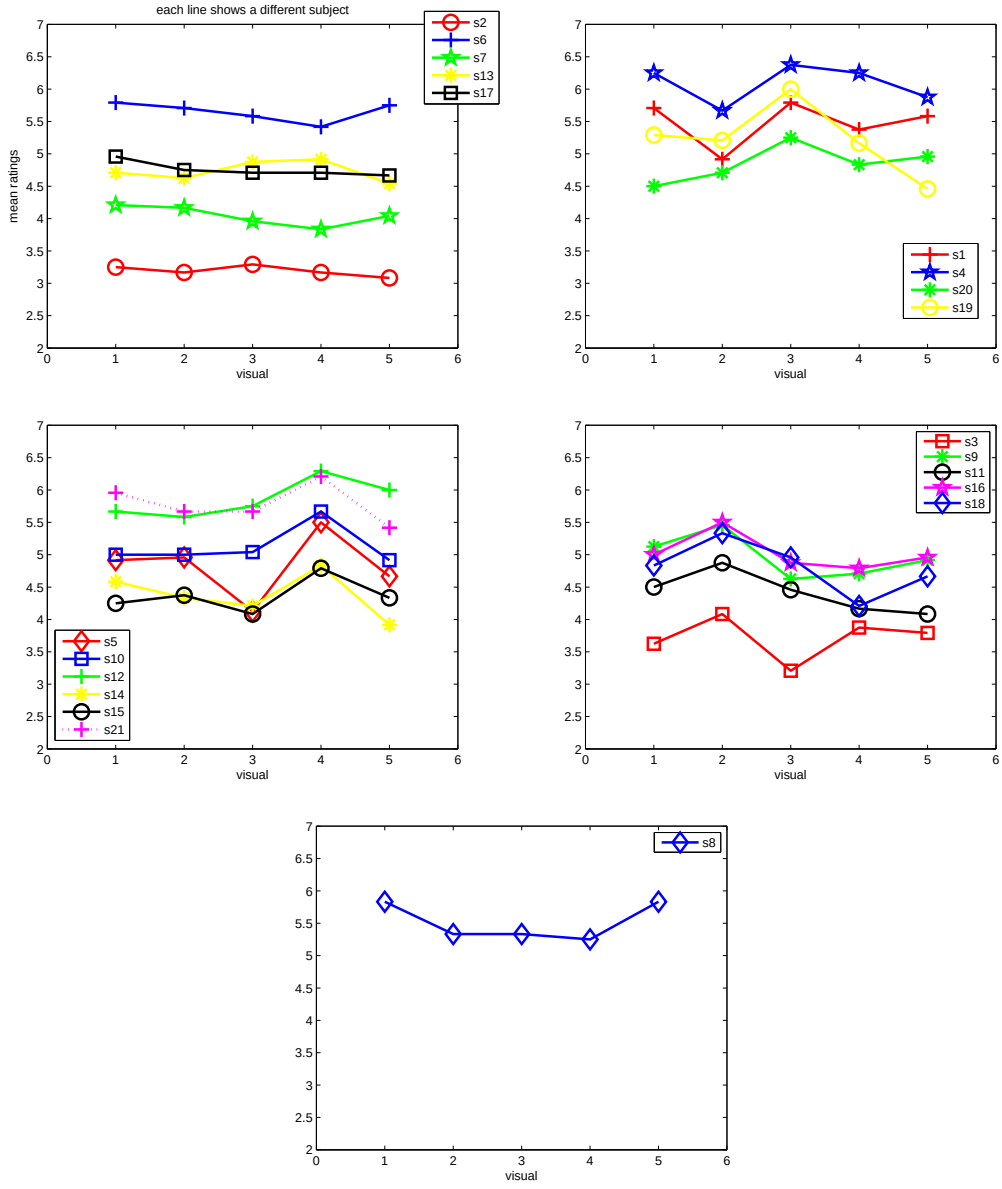


Fig. 12: Plots of means for the $S*V$ interaction (AV data) where the individual response of each subject is shown. Subject responses that are similar are grouped together in subplots. The `top left` plot features responses that have a slope almost equal to 0, although overall some subject's ratings are significantly higher than others. The `top right` plot features responses where visual stimuli 3 ($v3$) is the highest in quality. The `middle left` plot shows responses where $v4$ is the highest in quality. The `middle right` plot shows responses where $v2$ is the highest in quality. The `bottom` plot shows a response where stimuli $v1$ and $v5$ are the highest in quality, with the rest being almost equal.

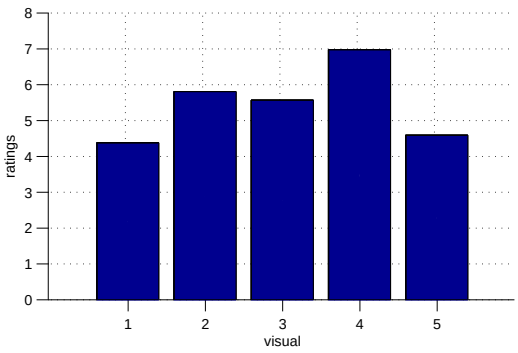


Fig. 15: Average ratings across subjects for the V-only experiment.

Analysis of Variance

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
S	169.67	20	8.4833	19.79	0
V	182.5	4	45.6262	106.46	0
S*V	625.1	80	7.8137	18.23	0
Error	45	105	0.4286		
Total	1022.27	209			

Constrained (Type III) sums of squares.

Fig. 16: The ANOVA table for the V-only experiment.

The average results across subjects for all visual-only presentations are shown in figure 15. The figure shows that the are large level differences. Visual stimulus 4 (*v4*) is the largest in size loudspeaker and rated highest, *v2* and *v3* are intermediate size loudspeakers, while *v1* and *v5* are the smallest in size loudspeakers and rated lowest. The ANOVA table is shown in figure 16. Factors *subjects* and *visual* as well as the 2-way interaction are all strongly significant.

3.4. Data across experiments

The overall ranking of audio and visual stimuli for the AV as well as the A-only and V-only experiments are shown in table 2. The overall rankings show close resemblance between audio ratings from the audiovisual and A-only experiments while the visual rankings are different between experiments. Figures 17 and 18 show the means ± standard deviations of the data. Close resemblance is seen for the audio ratings, while mean visual ratings are different between the AV and V-only experiments.

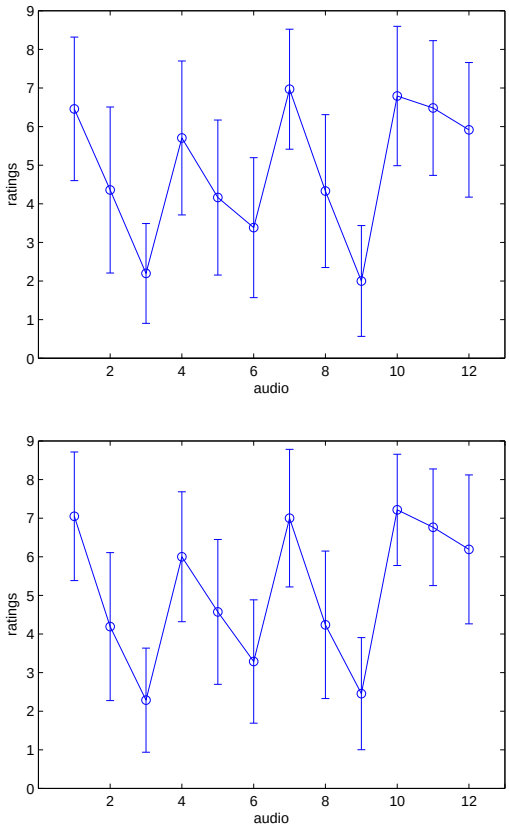


Fig. 17: Plots of the average ratings ± 1 standard deviation for the audio stimuli in the AV (top plot) and A-only (bottom plot) experiments.

The ANOVA between the AV experiment and the A-only, V-only experiments is shown in figures 19 and 20. In each ANOVA the 2 experiments are modeled by factor *experiment* (Exp). The figures show that factor *Exp* is statistically significant in both cases. This result indicates that the data obtained in the A-only and V-only experiments are somewhat different to the data in the AV experiment. Furthermore, the $V * Exp$ interaction is statistically significant while $A * Exp$ is statistically not significant. The non-significance in the $A * Exp$ interaction indicates that the ratings for each level of factor *audio* were similar between the 2 experiments.

Table 2: Rankings across experiments, averaged across subjects. The ranking order is shown from lowest to highest. A(AV) and V(AV) are the averaged results for the audiovisual experiment with respect to the audio and visual stimuli respectively.

experiment	A(AV)	V(AV)	A-only	V-only
ranking	3,6,5,2,4,1	5,3,2,4,1	3,6,2,5,4,1	1,5,3,2,4

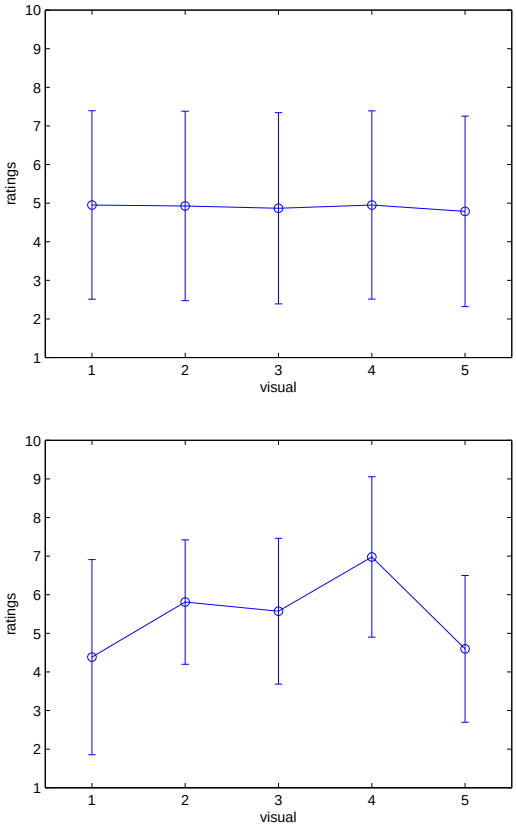
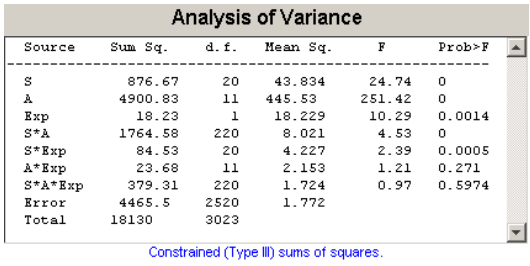


Fig. 18: Plots of the average ratings ± 1 standard deviation for the visual stimuli in the AV (top plot) and V-only (bottom plot) experiments.

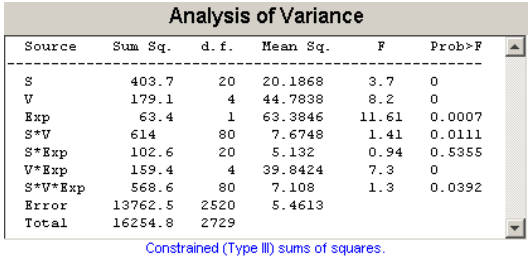


Analysis of Variance

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
S	876.67	20	43.834	24.74	0
A	4900.83	11	445.53	251.42	0
Exp	18.23	1	18.229	10.29	0.0014
S*A	1764.58	220	8.021	4.53	0
S*Exp	84.53	20	4.227	2.39	0.0005
A*Exp	23.68	11	2.153	1.21	0.271
S*A*Exp	379.31	220	1.724	0.97	0.5974
Error	4465.5	2520	1.772		
Total	18130	3023			

Constrained (Type III) sums of squares.

Fig. 19: ANOVA between the audio data in the AV and A-only experiments.



Analysis of Variance

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
S	403.7	20	20.1868	3.7	0
V	179.1	4	44.7838	8.2	0
Exp	63.4	1	63.3846	11.61	0.0007
S*V	614	80	7.6748	1.41	0.0111
S*Exp	102.6	20	5.132	0.94	0.5355
V*Exp	159.4	4	39.8424	7.3	0
S*V*Exp	568.6	80	7.108	1.3	0.0392
Error	13762.5	2520	5.4613		
Total	16254.8	2729			

Constrained (Type III) sums of squares.

Fig. 20: ANOVA between the visual data in the AV and V-only experiments.

4. CONCLUSION

The main goal of the experiments described in this paper was to investigate the influence of the visual appearance of products on audio quality evaluation. If the results of the audiovisual experiment could be explained and predicted by the results of the separate audio and visual experiments, then the audiovisual experiment would not be required. This would effectively imply that modality interactions are not important, and from a product design perspective that would mean that the modalities are independent so that for example improvements in one modality with the other modality unaltered would have an overall positive effect. This paper shows that the dominance of the auditory modality is not predicted by the unimodal evaluations.

4.1. Main factors in the AV, A-only and V-only experiments

The analysis for both the AV and A-only experiments shows factor *audio* to be the term with the largest effect. The similarity between the AV and A-only ratings shows that factor *audio* dominates the audiovisual subjective evaluation. In the AV experiment, factor *visual* was shown to have only a small influence, however the ANOVA shows that it is required to the overall statistical model and thus not unimportant as a source of information. In contrast to the AV results, the results of the V-only experiment show that factor *visual* is statistically significant. This indicates that when presented in isolation, the differences between the visual stimuli are perceived clearly and are judged to be substantial but become obscure in the presence of audio stimuli. A plausible explanation is that in the context of this experiment and due to the particularity of the product under test, audio has a larger impact and is a more decisive factor for the product's overall performance.

4.2. Interactions in the AV experiment

The plots showing the A-only and V-only ratings (figures 13 and 15) illustrate the perceived quality difference among the levels of each modality. According to [24] the quality balance between modalities is an important factor that can influence audiovisual evaluations. However, the V-only results show that there are large visual level differences that are comparable to the audio level differences. This indicates that the dominance of audio over visual constitutes a significant interaction.

Analysis of Variance

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
Sex	24.1	1	24.143	7.52	0.0061
A	7119.2	11	647.197	201.61	0
V	10.3	4	2.587	0.81	0.5213
Sex*A	110.8	11	10.069	3.14	0.0003
Sex*V	17.5	4	4.383	1.37	0.2435
A*V	142.6	44	3.242	1.01	0.4544
Sex*A*V	98.4	44	2.237	0.7	0.9351
Error	7704.5	2400	3.21		
Total	15169.1	2519			

Constrained (Type III) sums of squares.

Fig. 21: Analysis of the AV data including factor *Sex*.

4.3. Sex differences in audiovisual evaluation

According to popular belief men and women do not share the same appreciation for loudspeakers. An additional analysis on the results of the audiovisual experiment presented in this paper, including the subject's sex as a factor is shown in figure 21. The ANOVA shows that overall the responses by male and female subjects are different. The interaction *Sex*A* is statistically significant showing that there are differences in the way men and women evaluate audio quality, however the interaction *Sex*V* shows that there is no significant difference in the visual ratings by the 2 sex groups.

5. ACKNOWLEDGEMENTS

The authors would like to thank Claus Vestergaard Skipper and Peter Dissing for their help in the setup of the experiment and Søren Bech (Bang & Olufsen) for providing some of the loudspeakers used in the experiments.

6. REFERENCES

[1] Kohlrausch, A. and van de Par, S., "Audio-visual interaction in the context of multimedia applications", in J. Blauert (Ed.), "Communication Acoustics", Springer, Berlin, Germany, 2005, pp.109-138.

[2] ITU-R Rec. BS.1286, "Methods for the subjective assessment of audio systems with accompanying picture", International Telecommunications Union, Geneva, Switzerland, 1997.

[3] Thiede, T., Treurniet, W. C., Bitto, R., Schmidmer, C., Sporer, T., Beerends, J. G., Colomes, C., "PEAQ - The ITU Standard for Objective Measurement of Perceived Audio Quality", J. Audio Eng. Soc., Vol. 48, No. 1/2, 2000.

[4] Rix, A. W., Hollier, M.P, Hekstra A.P, Beerends, J.G., "Perceptual Evaluation of Speech Quality (PESQ). The New ITU Standard for End-to-End Speech Quality Assessment Part I – Time-Delay Compensation", J. Audio Eng. Soc., Vol. 50, No. 10, 2002.

[5] Takahashi, A., Hands, D. and Barriac, V., "Standardization Activities in the ITU for a QoE Assessment of IPTV", IEEE Communications Magazine, February 2008.

[6] Puria, A., Chen, X. and Luthra, A., "A Distortion Measure for Blocking Artefacts in Images Based on Human Visual Sensitivity", IEEE Transactions on Image Processing, **4**(6), 1995.

[7] Winkler, S., "Digital Video Quality: Vision Models and Metrics", Wiley, 2005.

[8] ITU-T Rec. P.910, "Subjective Video Quality Assessment Methods for Multimedia Applications". International Telecommunications Union, Geneva, Switzerland, 2008.

[9] ITU-R Rec. BT.500-12, "Methodology for the Subjective Assessment of the Quality of Television Pictures", International Telecommunications Union, Geneva, Switzerland, 2009.

[10] ITU-T Rec. J.144 , "Objective Perceptual Video Quality Measurement Techniques for Digital Cable Television in the Presence of a Full Reference", International Telecommunications Union, Geneva, Switzerland, 2004.

[11] ITU-R Rec. BT.1683, "Objective Perceptual Video Quality Measurement Techniques for Standard Definition Digital Broadcast Television in the Presence of a Full Reference", International Telecommunications Union, Geneva, Switzerland, 2004.

[12] Woszczyk, W., Bech, S., Hansen, V., "Interactions between audio-visual factors in a home theater system: Definition of subjective attributes", 99^{th} Audio Eng. Soc. Conv., New York, Preprint 4133:K-6, 1995.

[13] Insko, B.E. "Measuring Presence: Subjective, Behavioral and Physiological Methods", in Riva, G., Davide, F. and Ijsselsteijn, W.A., (Eds.), "Being There: Concepts, effects and measurement of user presence in synthetic environments", Ios Press, 2003.

[14] ITU-T Rec. J.147, "Objective picture quality measurement method by use of in-service test signals", International Telecommunications Union, Geneva, Switzerland, 2002.

[15] ITU-T Rec. P.911, "Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems", International Telecommunications Union, Geneva, Switzerland, 1998.

[16] ITU-R Rec. BS.1534-1, "Method for the subjective assessment of intermediate quality level of coding systems", International Telecommunications Union, Geneva, Switzerland, 2003.

[17] EBU Sound Quality Assessment Material CD, European Broadcasting Union, Geneva, Switzerland, 2008, Track 70: Eddie Rabbitt - "Early in the morning", timing [min.]: 0:00 - 0:09.

[18] Bob Marley and the Wailers, "Uprising - Coming In From The Cold", Island, 2001, ASIN: B00005A7X0, 44100 Hz, 16 bit, timing [min.]: 0:15 - 0:26.

[19] Can, "Soundtracts - She brings the rain", Mute U.S., 1998, ASIN: B0000067X8, 44100 Hz, 16bit, timing [min.]: 0:08 - 0:19.

[20] ITU-R Rec. BS.1116, "Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems", International Telecommunications Union, Geneva, Switzerland, 1997.

[21] Spence, C., "Audiovisual multisensory integration", Acoust. Sci. & Tech. 28, 2, 61-70, 2007.

[22] ITU-T Rec. J.140, "Subjective picture quality assessment for digital cable television systems", International Telecommunications Union, Geneva, Switzerland, 1998.

[23] Montgomery D., "Design and Analysis of Experiments", 5^{th}ed., Wiley, 2001, page 234.

[24] Hollier, M.P, and Voelcker, R.M., "Towards a multi-modal perceptual model", BT Technol. J. Vol. 14 No. 4 October 1997.

Manuscript B

Audiovisual subjective evaluation with an accompanying large-scale photograph

Alex Karandreas[1], Flemming Christensen[1]

[1] *Department of Electronic Systems, Acoustics, Aalborg University, DK-9220 Aalborg, Denmark*

Correspondence should be addressed to Alex Karandreas (aka@es.aau.dk)

ABSTRACT

The subjective audiovisual evaluation of loudspeakers was investigated. Loudspeaker reproduction was combined with visual presentations of 1:1 scale loudspeaker photographs. Additional unimodal experiments produced a baseline for comparison. Results indicate a stronger influence of the audio stimuli to the overall audiovisual evaluation and suggest that in audiovisual subjective evaluations a photograph presentation can be a valid substitute of the actual product.

1. INTRODUCTION

For most products a user's overall perception depends on more than one modality. For these reason, multisensory experiments are useful in the study of the influence of each modality. However, in the growing field of multisensory experiments there is lack of agreed upon practises [1]. The authors have previously designed audiovisual experiments [2] in an attempt to create a useful procedure for testing the overall audiovisual quality of loudspeakers. The experiment described here uses that procedure and the product under test are again loudspeakers, however the stimuli presentation is different. Instead of an actual product that produces all stimuli, in the experiment presented here a visual substitute - as close as possible to the real object - is used. Thus the aim of this study is to investigate the relative importance of the auditory and visual modalities to the overall quality evaluation as well as the validity of an alternative audiovisual stimuli presentation.

2. METHOD

2.1. Experimental design

The current experiment featured the following parts in order of presentation: audition and vision screening., familiarization, audiovisual experiment (simultaneous presentation of audio and visual stimuli), audio-only experiment (no visual stimuli) and visual-only experiment (no audio stimuli). The presentation of the two latter experiments was counterbalanced.

The experimental design was a full factorial design with absolute categorical scaling. Hence, all audio stimuli were combined with all visual stimuli and each combination was presented 4 times to each subject. The order of presentations was randomized and counter-balanced across subjects. The subjects task was to judge the audiovisual presentation and select an answer from a rating scale. The experimental question was: *"How does this loudspeaker sound?"* and the anchors were *"bad"* and *"excellent"*. A discrete 9 point rating scale was used. The rating scale and anchors was inspired by ITU recommendations [3] and [4]. A screenshot of the question and rating scale is shown in figure 1. The purpose of this experimental question was to urge subjects to consider the sound as an integral part of the product. Since the aim of the audiovisual experiment was to investigate the relative importance of the auditory and visual modalities to the overall quality evaluation, it was crucial that the subjects perceived the audiovisual presentation as an entity, so that the audio stimuli would not be perceived as an isolated phenomenon but an integral part of the loudspeaker presentation.

For the ratings of the audio-only experiment the experimental question was kept identical. For the ratings of the visual-only experiment the rating question was: *"How would this loudspeaker sound?"*. This question is similar to that of the audiovisual

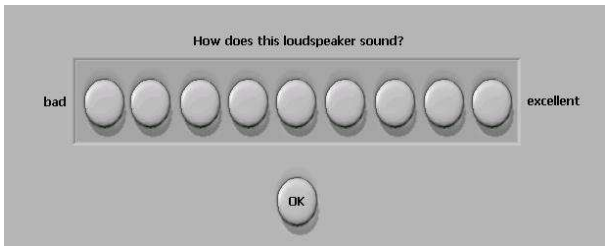


Fig. 1: Screenshot of the question and rating scale.

experiment, bearing in mind that there was no audio during this experiment.

Note that during the audio-only and the visual-only experiments the subjects were allowed (but not instructed) to give answers influenced by the previous audiovisual experiment (since they might have had associated an audio stimuli with a visual stimuli).

For the audio-only and visual-only experiments that followed the audiovisual experiment, each subject was presented 4 times with all stimuli. The presentations in each experiment were randomized and counter-balanced and the order of the 2 experiments was also counter-balanced.

2.2. Stimuli

The experiment featured 6 degraded versions of a single music excerpt. The excerpt features the chorus of a rock/country recording with male vocals, strumming acoustic guitar, snare drum, bass and handclaps. The music excerpt was selected from a reference recording [5] and transferred (ripped) to a computer (44.1kHz, 16bit). The excerpt included a complete musical phrase lasting 9 sec. There were 3 high-pass filtered versions and 3 with added harmonic distortion. The high-pass filtered versions were filtered at 110, 220 and 440 Hz while the harmonically distorted versions were all high-pass filtered at 110 Hz and had added harmonic distortion at 3 distinct levels. The pattern of harmonic distortion (2^{nd}, 3^{rd}, 4^{th} and 5^{th} order harmonics) was constant and the only difference was the relative level of the harmonic distortion to the 110 Hz high-pass filtered excerpt. The excerpts were not equalized with respect to loudness.

The same stimuli were successfully used in previous experiments, and were shown to be consistently ranked by a similar group of subjects [2].

In a pilot experiment [2] 12 loudspeaker models were evaluated and 5 models were judged to be quite different from one another. These 5 loudspeakers models were presented in this experiment as photographs (see figure 2) projected in 1:1 scale. The selected loudspeakers were:

1. Satellite (of a surround system) 1-way unit in grey plastic cabinet. Dimensions: 12.5 x 9 cm. Diaphragm not visible.
2. Large bookshelf 3-way loudspeaker with 4:3 cabinet proportions. Dimensions: 29 x 41 cm. Diaphragm not visible.
3. Large bookshelf 2-way unit with a rectangular wooden cabinet. Dimensions: 35 x 23 cm. Diaphragm visible.
4. Floor standing 4-way loudspeaker. Dimensions: 184 x 18.5 cm. Diaphragm not visible.
5. Small bookshelf 1-way unit in black plastic cabinet with a tilted upper section. Dimensions: 20.5 x 13 cm. Diaphragm not visible.

2.3. Setup

Excerpts were presented at comfortable listening levels through a single loudspeaker (Genelec $1031-A$). The loudspeaker was calibrated to produce a 75dB SPL at the listening position when reproducing 1/3 octave band-limited pink noise with center frequency either at 400 or 1000 Hz (-6dBFS at 44100 Hz, 16 bit). The loudspeaker was placed on the listening axis, and hidden behind a fake wall. A view of the setup can be seen in figure 3.

A monophonic playback system was chosen as a simple and valid approach in order to maintain a constant frequency and directivity response for all audio stimuli with a common on-axis path to the listener.

In order to ensure an unobstructed acoustical path from the loudspeaker to the listening position, the visual stimuli were projected onto 2 projector screens (Projecta, HomeScreen, 240x180cm, 4:3 aspect ratio, Matte White, reflection value 1, viewing angle 50 degrees L/R) that were separated with a 1 meter gap. A projector (Epson EMP-710) (see table 1) was placed within a sealed double window construction, behind the listening position. For each

Fig. 2: Photographs of loudspeakers used in this study: from left to right the visual stimuli are 1, 2, 3, 4 and 5.

visual stimulus, a custom-made digital picture was composed of 2 photographs of the loudspeaker separated by a vertical black strip running along the middle of the picture. When this image was projected onto the projector screens, the accurate dimensions of the loudspeakers were displayed on each projector screen while the black strip ensured that there was minimal light projected at the center area between the 2 screens and that could be reflected by the front wall causing glare.

The quality of the projection was tested with the NEC Test Pattern Generator 1.0 software tool. The noise of the projector at the listening position was measured to be below 35 dB SPL at all frequencies when the projector fan was idle. In order to avoid the fan turning on (the projector was placed in a small and airtight area) each session was limited to 8 minutes.

The rating scale was shown on a touch screen immediately after the stimuli presentation. During the presentation the screen was turned to black.

The listening room was kept completely dark. The light and viewing conditions in the laboratory room are shown in table 2.

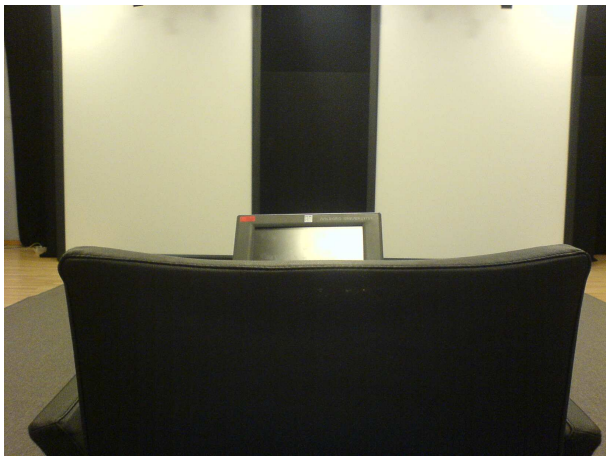


Fig. 3: Photograph of the setup. The 2 projector screens are visible in the background, and in the foreground the subject's chair and the touch screen.

2.4. Screening

3 university students (1 male and 2 female) participated in this study (mean age = 23 yrs, std = 1.7). An audiometric screening was made, and none had a hearing loss greater than 15dB HL in either ear at any octave band frequency between 125 Hz and 8 kHz. The participants were required to have normal or corrected to normal vision acuity and nor-

Table 1: Projector specifications according to the manufacturer

Aspect Ratio	4:3 (Native)
Native Resolution	1024 x 768
Brightness	1000 ANSI Lumens
Contrast Ratio	400:1
Operating Noise level	$<$ 40dB

Table 2: Light and viewing conditions in the laboratory room. Luminance measured with Gossen MAVOLUX 5032C, Class C acc. DIN 5032-7, Min. Sensitivity: $0.1lx$.

Background room illumination	0 lx
Listening room dimensions	$8.1 * 7.34 * 2.86$ m
Viewing distance to the touch screen	0.40 m
Viewing distance to the projector screens	6 m
Distance from projector to projector screen	7.5 m

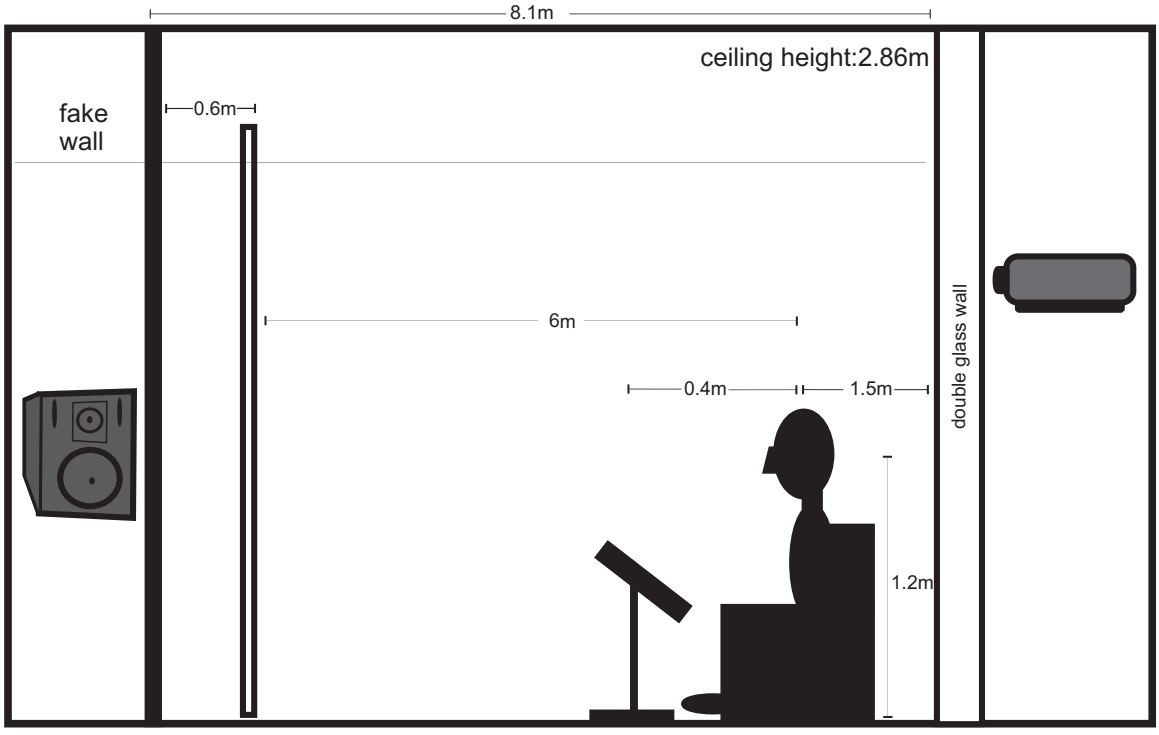


Fig. 5: Vertical view of the setup.

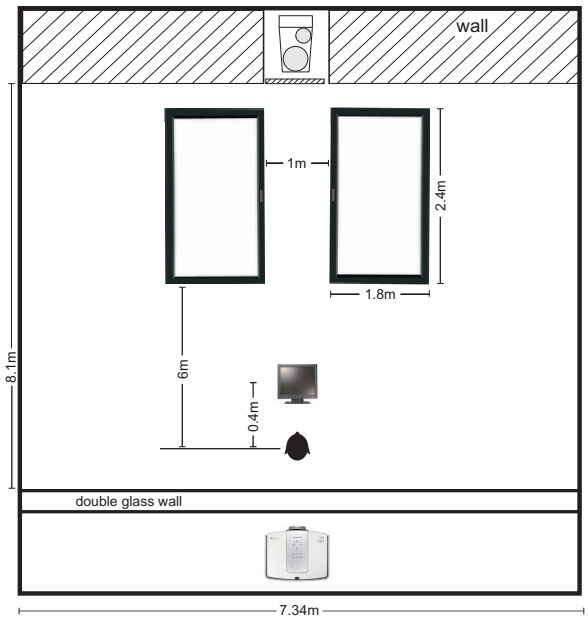


Fig. 4: Diagram of the setup.

mal color vision in accordance to ITU recommendations [3], [4], [6], [7]. The participants vision was inspected prior to the experiments using standard vision charts. Concerning acuity, no error on the 20/30 line of the standard eye chart was made. Concerning color vision, no more than 2 plates were missed out of 12 on an Ishihara test.

As a further supplement, prior to the experiment, data regarding the listening habits and prior experience of the participants were collected by means of a questionnaire to ensure that the subjects were naive listeners[1].

2.5. Familiarization

Prior to the experiment, subjects were introduced to the stimuli and the experiment's procedure. In 2 different familiarization sessions all visual and then all audio stimuli were presented in isolation. A third session featured selected audiovisual combinations presented in the same way as in the actual experiment.

[1] In the context of this paper, a naive subject is one that has no prior experience from viewing or listening experiments and has limited knowledge of loudspeakers (technical and commercial).

More precisely, during the first familiarization the visual stimuli were presented to the subjects without any reference to the rating scale or anchors. The second familiarization, introduced the range of audio degradations, presenting the least degraded and the most degraded stimuli. In order to present subjects with the range of audio degradations, only during this familiarization the audio stimuli were labeled: the least degraded stimulus was termed "excellent" and the most degraded stimuli (of both degradation methods) were termed "bad". Subjects were instructed that these stimuli were just some of the sounds they would hear during the experiment.

A third familiarization emulated the stimuli presentation and the evaluation procedure as it would take place during the actual experiment. Three audiovisual presentations were given, during which the least degraded stimulus, and the most degraded stimuli (of both degradation methods) were coupled with a loudspeaker. The loudspeaker presentation in this familiarization was also counter-balanced among subjects.

3. RESULTS

3.1. Audiovisual experiment results

The average results across subjects for all audiovisual combinations are shown in figure 6. The figure shows large differences (up to 4.5 rating points) between audio levels and smaller differences between visual levels (up to 1 rating point in most cases).

The ANOVA table is shown in figure 7. Factors *subjects*, *audio* and the 2-way interaction *subjects*audio* are significant. Factor *visual* is almost statistically significant however the Sums of Squares column shows that it is much less influential than the other main factors.

The ratio of the Sum of Squares of each term to the total Sum of Squares can be expressed as a percentage contribution of each term to the ANOVA model [8]. For this experiment, factor *audio* accounts for 42% of the variability of the experiment, while factor *visual* accounts for 0,95%.

The *subjects*audio* interaction plot (figure 8) shows large differences between the levels of factor *audio*. Furthermore, there are differences in the way each

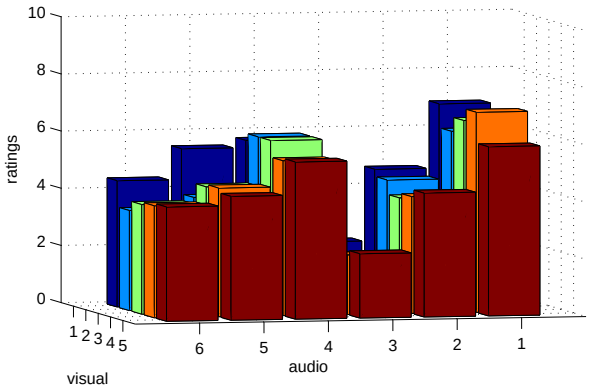


Fig. 6: Average ratings across subjects for all combinations of audio and visual stimuli, in the audiovisual experiment. A rating equal to 1 corresponds to low quality and 9 to high quality. Audio stimuli 1, 2 and 3 are high-pass filtered stimuli and stimuli 4, 5 and 6 are harmonically distorted stimuli.

Analysis of Variance

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
S	92.57	2	46.286	25.4	0
A	713.56	5	142.711	78.32	0
V	15.91	4	3.976	2.18	0.0713
S*A	263.56	10	26.356	14.46	0
S*V	19.76	8	2.47	1.36	0.2163
A*V	26.19	20	1.31	0.72	0.806
S*A*V	50.44	40	1.261	0.69	0.92
Error	492	270	1.822		
Total	1673.99	359			

Constrained (Type III) sums of squares.

Fig. 7: The ANOVA table for the audiovisual experiment, including 2 and 3-way interactions. S, A and V refer to factors *subjects*, *audio* and *visual* respectively.

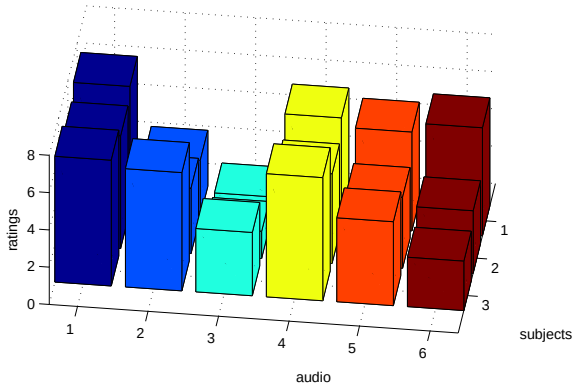


Fig. 8: Plot of the 2-way interaction between factors *subject* and *audio* for the audiovisual experiment.

subject evaluates the stimuli although all subjects follow the same trend, with the exception that subject 1 (*s1*) rates audio levels 4 (*a4*), *a5* and *a6* almost equally with *a6* rated higher than *a5*. The *subjects*visual* plot (figure 9) shows large differences between subjects, however they all follow the same trend, rating almost equally all visual stimuli with the exception of a large difference between interactions *s1*v1* and *s1*v2*.

Statistical analysis of the data shows that it deviates slightly from a normal distribution. Non-parametric analysis on the raw data shows results to be very similar to those of the ANOVA (the same factors are statistically significant).

3.2. Audio experiment results

The average results across subjects for all audio-only presentations are shown in figure 10. The figure shows that the are large differences between the ratings of the audio levels.

The ANOVA table is shown in figure 11. Factors *subjects* and *audio* are strongly significant, with *audio* the most influential term.

3.3. Visual experiment results

The average results across subjects for all visual-only presentations are shown in figure 12. The figure

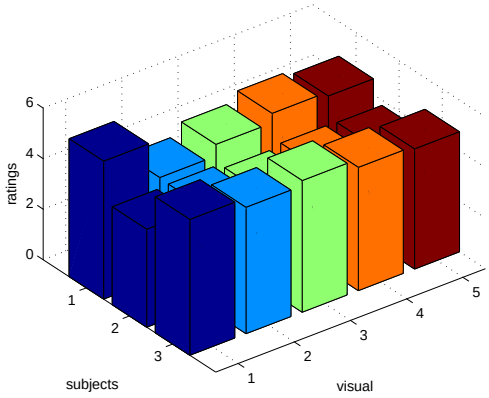


Fig. 9: Plot of the 2-way interaction between factors *subject* and *visual* for the audiovisual experiment.

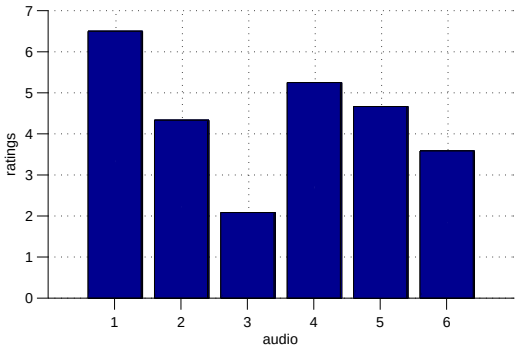


Fig. 10: Average ratings across subjects for the audio-only experiment.

Analysis of Variance

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
S	27.111	2	13.5556	6.64	0.0026
A	134.903	5	26.9806	13.21	0
S*A	31.056	10	3.1056	1.52	0.1573
Error	110.25	54	2.0417		
Total	303.319	71			

Constrained (Type III) sums of squares.

Fig. 11: The ANOVA table for the audio-only experiment.

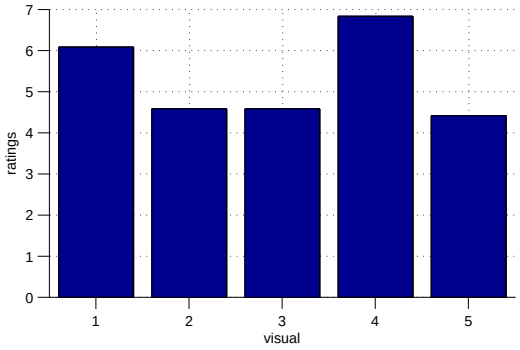


Fig. 12: Average ratings across subjects for the visual-only experiment.

Analysis of Variance

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
S	15.6	2	7.8	17.55	0
V	57.267	4	14.3167	32.21	0
S*V	137.733	8	17.2167	38.74	0
Error	20	45	0.4444		
Total	230.6	59			

Constrained (Type III) sums of squares.

Fig. 13: The ANOVA table for the visual-only experiment.

shows that visual stimuli 2 (*v2*), *v3* and *v5* are rated almost equally and stimuli *v1* and *v4* rated higher with a maximum difference between levels of about 2.5 points.

The ANOVA table is shown in figure 13. Factors *subjects*, *visual* and the 2-way interaction are all strongly significant.

3.4. Data across experiments

The overall ranking of audio and visual stimuli for the audiovisual as well as the audio-only and visual-only experiments are shown in table 3. The audio stimuli are identically ranked in the audiovisual and audio-only experiments, while the visual stimuli ranking between experiments is different. Figures 14 and 15 show the means ± standard deviations of the data in the 3 experiments. Close resemblance is seen for the audio ratings in the audiovisual and

Table 3: Data across experiments, averaged across subjects. The ranking order is shown from lowest to highest. A(AV) and V(AV) are the averaged results for the audiovisual experiment with respect to the audio and visual stimuli respectively.

experiment	A(AV)	V(AV)	Audio-only	Visual-only
ranking	3,6,2,5,4,1	2,5,3,4,1	3,6,2,5,4,1	5,3,2,1,4

audio-only experiments, while visual ratings are different between the audiovisual and visual-only experiments. For the mean visual rating in the visual-only experiment there is a difference between levels with visual stimuli 1 and 4 having the highest rating. In the audiovisual experiment there are hardly any differences between visual levels although stimulus 1 is rated highest.

3.5. CONCLUSIONS

The analysis for the audiovisual experiment and the comparison with the audio-only experiment shows that factor *audio* was the term with the largest effect and that it dominated the audiovisual subjective evaluation. In the audiovisual experiment, factor *visual* was shown to be nearly statistically significant but to have only a small influence. In contrast to the audiovisual results for factor *visual*, the results of the visual-only experiment show that factor *visual* was statistically significant with much greater level differences. This indicates that when presented in isolation, the differences between the visual stimuli are perceived more clearly but become obscure in the presence of audio stimuli. A plausible explanation is that in the context of this experiment with the products under test being loudspeakers, audio reproduction is the decisive factor for the product's overall performance.

The larger influence of the auditory over the visual modality could not have been predicted without a multimodal evaluation. Therefore the results presented in this study suggest that multimodal evaluation is a useful tool for product quality evaluations.

One aim of this study was to investigate whether the presence of the actual product is necessary in the experiment or whether substitutes can be a valid alternative. In a previous benchmark experiment [2] the actual loudspeakers were presented as visual stimuli,

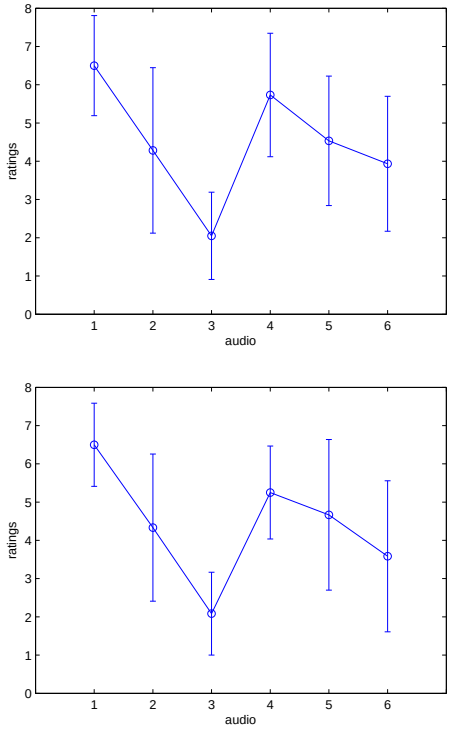


Fig. 14: Plots of the mean and ± 1 standard deviation for the audio stimuli of the audiovisual (top plot) and audio-only (bottom plot) experiments.

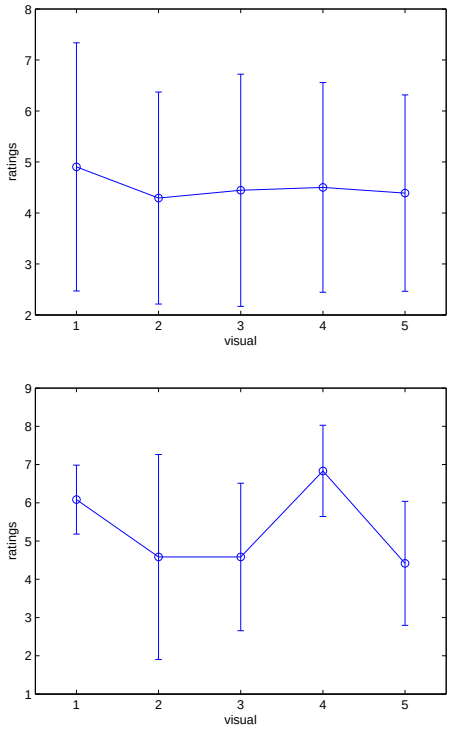


Fig. 15: Plots of the mean and ± 1 standard deviation for the visual stimuli of the audiovisual (top plot) and visual-only (bottom plot) experiments.

while the audio reproduction was identical. The 1:1 scale photographs are an attempt to a smooth transition from a 3D to a 2D presentation and the presentation of pairs of loudspeakers was not only due to practical considerations (having an unobstructed acoustical path to the listening position) but also a link to the presentation of pairs of actual loudspeakers in the benchmark experiment.

The results in this paper consist of data from only 3 subjects and they should therefore be considered with caution. The results are in agreement with the results of the benchmark experiment [2]. The ANOVA of the audiovisual, audio-only and visual-only experiment in this study and that of the benchmark study show that the factors have a similar effect. The only difference in the ANOVA analysis of the audiovisual experiments was that in the benchmark experiment the subjects*visual and audio*visual interactions were statistically significant. The ranking and ratings of the audio and visual stimuli are also similar and the same general conclusions can be drawn from both experiments. Thus, the results of this study indicate that under the given circumstances audiovisual experiments might be moved to simpler setups with the use of substitutes, but the risk exists that subtle modal interactions might be lost.

4. REFERENCES

[1] Kohlrausch, A. and van de Par, S., "Audio-visual interaction in the context of multimedia applications", in J. Blauert (Ed.), Communication Acoustics, Springer, Berlin, Germany, 2005, pp.109-138.

[2] Karandreas, A., Christensen, F., "Influence of visual appearance on loudspeaker sound quality evaluation". Submitted to the Journal of the Audio Engineering Society, 2010.

[3] ITU-T Rec. P.910, "Subjective Video Quality Assessment Methods for Multimedia Applications". International Telecommunications Union, Geneva, Switzerland, 2008.

[4] ITU-R Rec. BT.500-12, "Methodology for the Subjective Assessment of the Quality of Television Pictures", International Telecommunications Union, Geneva, Switzerland, 2009.

[5] EBU Sound Quality Assessment Material CD, European Broadcasting Union, Geneva, Switzerland, 2008, Track 70: Eddie Rabbitt - "Early in the morning", timing [min.]: 0:00 - 0:09.

[6] ITU-T Rec. P.911, "Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems", International Telecommunications Union, Geneva, Switzerland, 1998.

[7] ITU-T Rec. J.140, "Subjective picture quality assessment for digital cable television systems", International Telecommunications Union, Geneva, Switzerland, 1998.

[8] Montgomery D., "Design and Analysis of Experiments", 5^{th}ed., Wiley, 2001, page 234.

Manuscript C

Subjective assessment of loudspeaker reproduction with accompanying small-scale photograph - an audiovisual experiment.

Alex Karandreas[1], Flemming Christensen[1]

[1] *Department of Electronic Systems, Acoustics, Aalborg University, DK-9220 Aalborg, Denmark*

Correspondence should be addressed to Alex Karandreas (aka@es.aau.dk)

ABSTRACT

Loudspeaker reproduction of music excerpts complemented with the simultaneous presentation of loudspeaker photographs was evaluated in a subjective test. Additional unimodal experiments produced a baseline for comparison. Results indicate that the influence of audio stimuli dominates the audiovisual evaluation and that in audiovisual subjective evaluations the presentation of the actual product under evaluation can be substituted by a small-scale photograph.

1. INTRODUCTION

A user's perception of a product usually depends on more than one modality. Multimodal experiments can be used to study the influence of each modality. However, in the growing field of multimodal experiments there is lack of a common methodology [1] and a striking lack of literature on multimodal investigations of products. The authors have previously designed a benchmark audiovisual experiment [2] in an attempt to create a useful procedure for testing the overall quality of audiovisual products. That benchmark experiment featured actual loudspeakers as the source of both audio and visual stimuli. A point of concern with that study were the practical difficulties of the experimental setup. It was thus desired to extend the methodology to include product substitutes in an attempt to resolve these practical difficulties. This study investigates whether shifting from the actual loudspeaker (3D) visual presentation to a small scale 2D presentation (a small-scale photograph of the same product shown on a computer screen), will produce comparable results to those of the benchmark study. Thus, the experiment described here uses that same procedure as well as the same audio reproduction method, and the only difference lies in the presentation of the visual stimuli. Thus the aim of this study is to investigate the relative importance of audiovisual stimuli as well as the validity of an alternative audiovisual stimuli presentation. The present study evaluates overall impression in relation to audition and vision, using loudspeakers as an example.

2. METHOD

In order to enable a bimodal study of audiovisual interaction, both auditory and visual stimuli should be applied in a way that the test subjects are exposed to a range of different levels of the two modalities studied.

2.1. Stimuli

The experiment featured 6 degraded versions of a single music excerpt. The excerpt features the chorus of a rock/country recording with male vocals, strumming acoustic guitar, snare drum, bass and handclaps. The music excerpt was selected from a reference recording [3] and transferred (ripped) to a computer (44.1kHz, 16bit). The excerpt was carefully selected to include a complete musical phrase lasting 9 sec. There were 3 high-pass filtered versions and 3 with added harmonic distortion. The high-pass filtered versions were filtered at 110, 220 and 440 Hz while the harmonically distorted versions were all high-pass filtered at 110 Hz [1] and had

[1] Combining non-high-pass filtered audio stimuli with some of the loudspeakers would seem unnatural, as quite small loudspeakers would be coupled with audio stimuli that feature substantial energy at the low frequencies. Therefore, to

added harmonic distortion at 3 distinct levels. The pattern of harmonic distortion was constant and the only difference was the relative level of the harmonic distortion to the 110 Hz high-pass filtered excerpt.

The same stimuli were successfully used in previous experiments, and were shown to be consistently ranked by a similar group of subjects [2] [2].

In previous experiments [2], 5 different loudspeaker models were selected to be the visual stimuli and the actual loudspeakers were presented:

- Satellite (of a surround system) 1-way unit in grey plastic cabinet. Dimensions: 12.5 x 9 cm. Diaphragm not visible.
- Large bookshelf 3-way loudspeaker with 4:3 cabinet proportions. Dimensions: 29 x 41 cm. Diaphragm not visible.
- Large bookshelf 2-way unit with a rectangular wooden cabinet. Dimensions: 35 x 23 cm. Diaphragm visible.
- Floor standing 4-way loudspeaker. Dimensions: 184 x 18.5 cm. Diaphragm not visible.
- Small bookshelf 1-way unit in black plastic cabinet with a tilted upper section. Dimensions: 20.5 x 13 cm. Diaphragm not visible.

In this experiment instead of the actual loudspeakers, photographs (figure 1) were presented at a 12 inch monitor. The photographs were taken in a very controlled manner and they portrayed an accurate scale of the original. Audio quality evaluation with accompanying photograph presentation is considered to be a valid practise according to ITU recommendations [4].

2.2. Experimental design

The study featured the following parts in order of presentation: audition and vision screening, familiarization, audiovisual experiment (simultaneous presentation of audio and visual stimuli), audio-only experiment (no visual stimuli) and visual-only experiment (no audio stimuli). The presentation of the two latter experiments was counter-balanced.

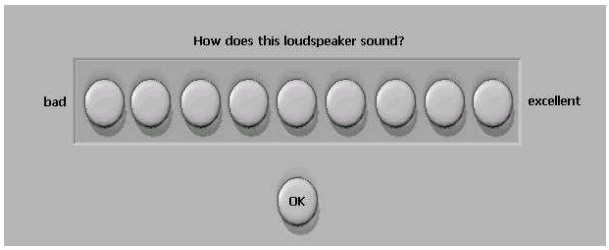


Fig. 2: Screenshot of the question and rating scale. The rating scale was shown once the audiovisual presentation was over. The subjects used the buttons to give their evaluation and then pressed the OK button to proceed to the next presentation.

A full factorial design with absolute categorical scaling was used. Hence, all audio stimuli were combined with all visual stimuli and each combination was presented 4 times to each subject. The order of presentations was randomized and counter-balanced across subjects. The subjects task was to evaluate each presentation using a rating scale with 9 discrete points. The rating scale and anchors were inspired by ITU recommendations [5] and [6]. For the audiovisual and audio-only experiments the experimental question was: *How does this loudspeaker sound?* and the anchors were *bad* and *excellent*. A screenshot of the question and rating scale is shown in figure 2. Since the aim of the audiovisual experiment was to determine the degree to which visual bias influences audio perception, it was crucial that the subjects perceived the audiovisual presentation as an entity. The purpose of this experimental question was to urge subjects to consider the sound as an integral part of the product.

For the visual-only experiment the rating question was altered to: *How would this loudspeaker sound?*. It was desired to maintain the question as similar as possible to that of the audiovisual experiment, bearing in mind that there was no audio during this experiment.

For the audio-only and visual-only experiments that followed the audiovisual experiment, each subject

ensure that the combination of audio stimuli and loudspeaker was not unnatural, the original excerpt was high-pass filtered at 110 Hz. All other degradations were made using these already filtered stimuli.

[2] A precise description of the stimuli and justification of their selection can be found in [2].

Fig. 1: The loudspeaker photographs featured in this study. From left to right the visual stimuli are 1, 2, 3, 4 and 5.

was presented 4 times with all stimuli. The presentations in each experiment were randomized and counter-balanced and the order of the 2 experiments was also counter-balanced. During the visual-only experiment each trial featured one photograph shown for 5 seconds. During the audiovisual experiment the photograph was shown for the exact duration of the audio stimuli. For the audio-only experiment the screen remained black during the audio presentation. For all experiments, after each trial the screen changed to show the rating scale.

Note that during the audio-only and visual-only experiments subjects were allowed (but not instructed) to give answers influenced by the previous audiovisual experiment (since they might have had associated an audio stimulus with a visual stimulus).

2.3. Screening

6 university students (3 male and 3 female) participated in this study (mean age = 24.6 yrs, std = 4.7). An audiometric screening was made, and none had a hearing loss greater than 15dB HL in either ear at any octave band frequency between 125 Hz and 8 kHz. The participants were required to have normal or corrected to normal vision acuity and normal color vision. The participants vision was inspected prior to the experiments using standard vision charts. Concerning acuity, no error on the 20/30 line of the standard eye chart was made. Concerning color, no plates were missed out of 12 on an Ishihara test (requirement was no more than 2 out of 12) [5], [6], [7], [8].

As a further supplement, prior to the experiment, data regarding the listening habits and prior experience of the participants were collected by means of a questionnaire to ensure that the subject were naive [3] listeners.

2.4. Familiarization

As mentioned in section 2.2, subjects were familiarized with the stimuli and the experimental procedure. In 2 different familiarization sessions all visual stimuli and the range of audio stimuli were presented in isolation. A third session featured selected audiovisual combinations presented in the same way as in the actual experiment.

More precisely, during the first familiarization the visual stimuli were presented to the subjects without any reference to the rating scale or anchors. The second familiarization, introduced the range of audio degradations, presenting the least and most degraded stimuli. In order to present subjects with the range of audio degradations, only during this familiarization the audio stimuli were labeled. The least degraded stimulus was termed *excellent* and the most degraded stimuli (of both degradation methods) were termed *bad*. Subjects were instructed that these stimuli were just some of the audio stimuli they would hear during the experiment.

A third familiarization emulated the stimuli presentation and the evaluation procedure as it would take

[3] In the context of this paper, a naive subject is one that has no prior experience from viewing or listening experiments and has limited knowledge of loudspeakers (technical and commercial).

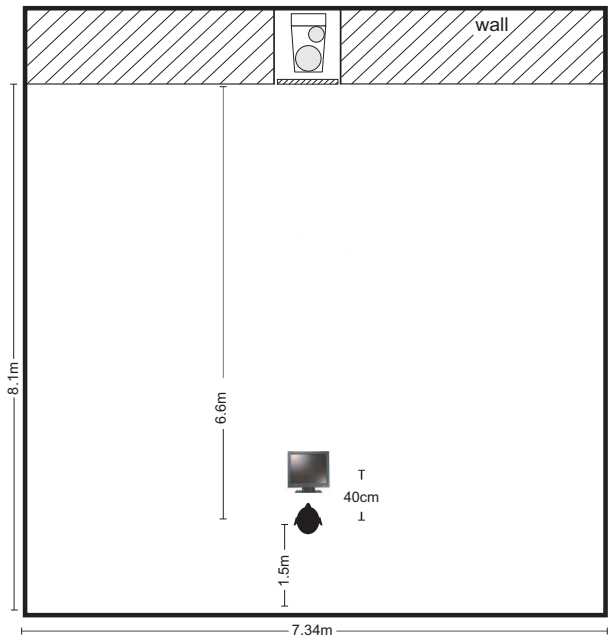


Fig. 3: Top view of the setup in the listening room.

place during the actual experiment. Three audiovisual presentations were given, during which the least degraded audio stimulus, and the two most degraded audio stimuli (high-pass filtering for one excerpt, harmonic distortion for the other) were coupled with a loudspeaker photograph. The presented loudspeaker was also counter-balanced among subjects.

2.4.1. Setup

The experiment was conducted in a laboratory room conforming with the ITU-R BS.1116 recommendation for listening rooms suitable for evaluations of multichannel audio systems [9], and controlled from an adjacent control room. The setup can be seen in figures 3 and 4.

The visual stimuli were shown on a touch screen (ELO Touchsystems ETL12IC, diagonal size: 12 inches) in front of the subjects. The size of the photographs displayed on the screen was 23 x 16.5 cm. The screen was adjusted on a stand at a height of 70 cm from the floor, an angle of 30^0 with respect to the floor and a viewing distance of approximately 40 cm.

The audio stimuli presentation was done by means of

Fig. 4: Photograph of the listening room showing the listening position and touch screen.

a single loudspeaker (Genelec 1031-A) on-axis to the listening position. This loudspeaker was mounted inside a fake wall, behind a fabric surface and was thus invisible to the subject. The geometrical center of the loudspeaker was at a height of 1.2 m. The direct path from loudspeaker to listening position was acoustically unobstructed. The loudspeaker was calibrated to produce a 75dB SPL at the listening position when reproducing 1/3 octave band-limited pink noise with center frequency either at 400 or 1000 Hz (-6dBFS at 44100 Hz, 16 bit). The stimuli were not equalized with respect to loudness.

The experiment was fully automated and controlled by a PC running custom-made software (programmed in Labview 6.1). This included the order of presentation of the stimuli (randomization), the stimuli generation and the data collection. The audio stimuli were generated by an internal sound card (RME Digi 9636, 24-bit, 96 kHz), were converted to analog (Tracer Technologies Big Daddi, 24 bit D/A converter) and reproduced by an active loudspeaker (Genelec 1031-A, Free field frequency response: 47-22000 Hz ($\pm$ 3dB), crossover frequency: 2.2 kHz, short term RMS @ 0.5m $>$ 107 dB SPL).

In order to have a constant frequency and directivity response for all audio stimuli, with a common on-axis path to the listener and keep the influence of room reflections as constant as possible, a monophonic playback was chosen since it presents a simple and valid approach.

Table 1: Light and viewing conditions in the laboratory room. Luminance measured with Gossen MAVOLUX 5032C, Class C acc. DIN 5032-7, Min. Sensitivity: $0.1lx$.

Background room illumination	0 lx
Listening room dimensions	$8.1 * 7.34 * 2.86$ m
Viewing distance to the touch screen	0.40 m

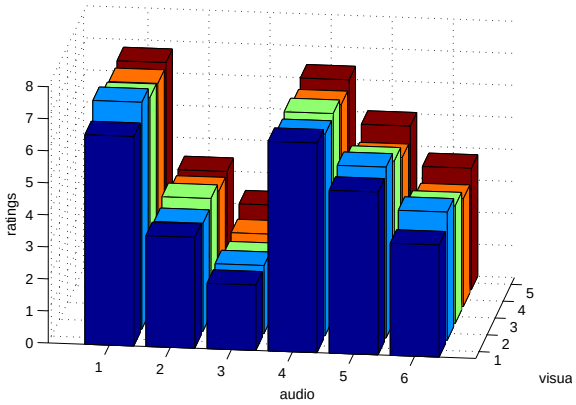


Fig. 5: Average ratings across subjects for all audiovisual presentations. A rating equal to 1 corresponds to low quality and 9 to high quality. Audio stimuli 1, 2 and 3 are high-pass filtered stimuli and stimuli 4, 5 and 6 are harmonically distorted stimuli.

Analysis of Variance

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
S	156.66	5	31.332	20.01	0
A	1958.98	5	391.796	250.16	0
V	10.02	4	2.505	1.6	0.1731
S*A	504	25	20.16	12.87	0
S*V	29.33	20	1.467	0.94	0.5404
A*V	13.85	20	0.692	0.44	0.9839
S*A*V	101.3	100	1.013	0.65	0.996
Error	845.75	540	1.566		
Total	3619.89	719			

Constrained (Type III) sums of squares.

Fig. 6: The ANOVA table for the audiovisual data including 2 and 3-way interactions. S, A and V refer to factors *subjects*, *audio* and *visual* respectively.

The listening room was kept completely dark. The light and viewing conditions in the laboratory room are shown in table 1.

3. RESULTS

3.1. Audiovisual experiment results

The average results across subjects for all audiovisual presentations are shown in figure 5. Large differences are seen on the ratings along the audio axis whereas the ratings along the visual axis show only small differences.

The ANOVA table is shown in figure 6. Factors *subjects*, *audio* and the 2-way interaction *subjects*audio* are statistically significant. The ANOVA analysis suggests that factor *visual* has negligible influence to the results. The ratio of the Sum of Squares of each factor to the total Sum of Squares can be expressed as a percentage contribution of each factor to the ANOVA model [10]. For this experiment, factor *audio* accounts for 54% of the variability of the experiment, while factor *visual* accounts for 0,3%.

The 2-way interactions can be viewed as plots in figures 5, 7 and 8. The individual $subjects * audio$ evaluations of each subject (figure 7) show a common pattern with subject 1 ($s1$) and $s4$ deviating from the general trend. More precisely, all responses show a highest rating for audio level 1 ($a1$), a decreasing rating for $a2$ and $a3$, an increase for $a4$ and a steady decrease for $a5$ and $a6$. The responses for $s1$ and $s4$ follow this trend but the slopes are more narrow. Subject $s4$ gives a relatively low rating for the two least degraded excerpts $a1$ and $a4$.

The individual $subjects * visual$ evaluations of each subject (figure 8) show almost identical responses by each subject for each visual level. More precisely, the response of each subject has small fluctuations but is otherwise nearly parallel to the x-axis. Furthermore, these lines are almost parallel to each other

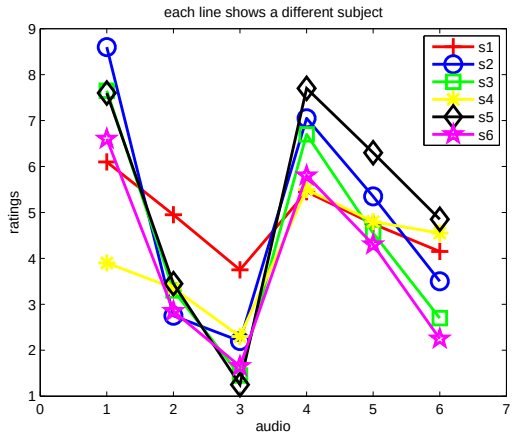


Fig. 7: Plot of the 2-way interaction between factors *subject* and *audio*, for the audiovisual data. The legend indicates the correspondence between subjects and responses.

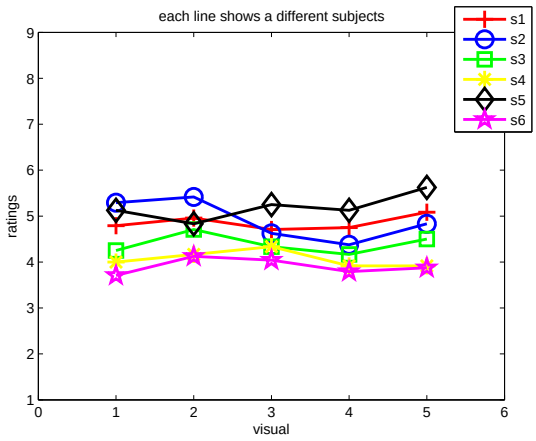


Fig. 8: Plot of the 2 way interaction between factors *subject* and *visual*, for the audiovisual data. The legend indicates the correspondence between subjects and responses.

and are all within a small range of the rating scale. This pattern of response is in agreement with the ANOVA table which shows a non-significant *subject*visual* interaction.

Statistical analysis of the data shows that it is not normally distributed, but rather skewed with a longer tail on the right side (that is on the higher quality ratings). This could be due to the discrete nature of the rating scale. Attempts to normalize the data had minimal effect and did not change the shape of the distribution. Non-parametric analysis on the raw data showed results to be very similar to those of the ANOVA (the same factors are statistically significant).

3.2. Audio experiment results

The average results across subjects for all audio-only presentations are shown in figure 9. The figure shows that the are large differences between the ratings of the high-pass filtered stimuli (*a1*, *a2* and *a3*) and between the harmonically distorted stimuli (*a4*, *a5* and *a6*).

The ANOVA table is shown in figure 10. Factors *subjects*, *audio* and the 2-way interaction are all strongly significant.

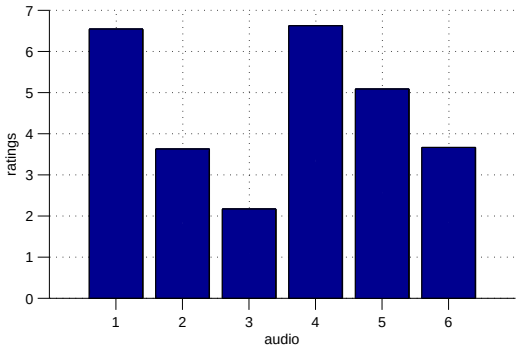


Fig. 9: Average ratings across subjects for the audio-only experiment.

Analysis of Variance

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
S	42.868	5	8.5736	5.49	0.0002
A	380.285	5	76.0569	48.68	0
S*A	128.09	25	5.1236	3.28	0
Error	168.75	108	1.5625		
Total	719.993	143			

Constrained (Type III) sums of squares.

Fig. 10: The ANOVA table for the audio-only data.

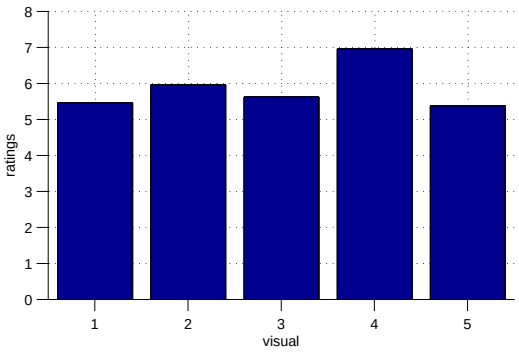


Fig. 11: Average ratings across subjects for the visual-only experiment.

Analysis of Variance

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
S	206.375	5	41.275	174.81	0
V	40	4	10	42.35	0
S*V	191.5	20	9.575	40.55	0
Error	21.25	90	0.2361		
Total	459.125	119			

Constrained (Type III) sums of squares.

Fig. 12: The ANOVA table for the visual-only data.

3.3. Visual experiment results

The average results across subjects for all visual-only presentations are shown in figure 11. The figure shows that the are small variations between the levels. The maximum difference between levels being about 2 rating points.

The ANOVA table is shown in figure 12. Factors *subjects*, *visual* and the 2-way interaction are all strongly significant. However, judging on the Sum of Squares terms, factor *subjects* and the 2-way interaction are much more influential than factor *visual*.

3.4. Data across experiments

The overall ranking of audio and visual stimuli for the audiovisual as well as the audio-only and visual-only experiments are shown in table 2. The audio stimuli are almost identically ranked in the audiovisual and audio-only experiments (there is an inversion between stimuli 1 and 4), while the visual stimuli ranking between experiments is different. Figures 13 and 14 show the means ± standard deviations of the data in the 3 experiments. Close resemblance is seen for the audio ratings in the audiovisual and audio-only experiments, while visual ratings are different between the audiovisual and visual-only experiments. In the audiovisual experiment there are hardly any differences between visual levels, while a level difference can be seen for the visual-only experiment.

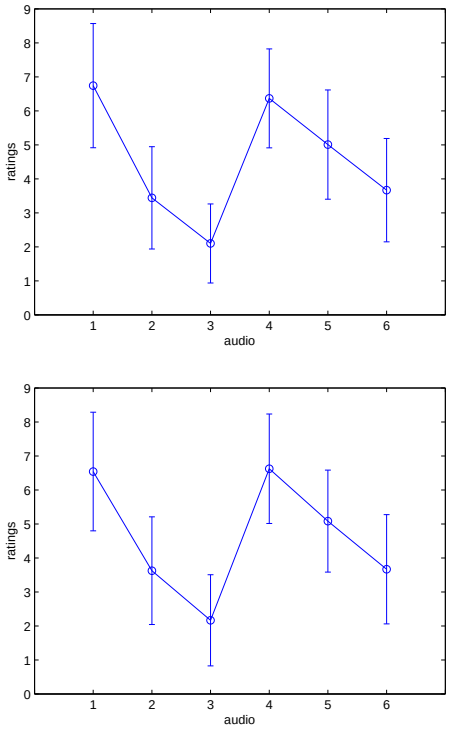


Fig. 13: Plots of the mean and ± 1 standard deviation for the audio stimuli of the audiovisual (top plot) and audio-only (bottom plot) experiments.

3.5. CONCLUSION

The analysis for the audiovisual and audio-only data shows factor *audio* to be the term with the largest effect. The similarity between the audiovisual and audio-only ratings shows that factor *audio* dominates the audiovisual subjective evaluation. In the audiovisual experiment, factor *visual* was shown to have only a small influence. In contrast to the audiovisual results, the results of the visual-only ex-

Table 2: Data across experiments, averaged across subjects. The ranking order is shown from lowest to highest. A(AV) and V(AV) are the averaged results for the audiovisual experiment with respect to the audio and visual stimuli respectively.

experiment	A(AV)	V(AV)	Audio-only	Visual-only
ranking	3,2,6,5,4,1	4,1,3,5,2	3,2,6,5,1,4	5,1,3,2,4

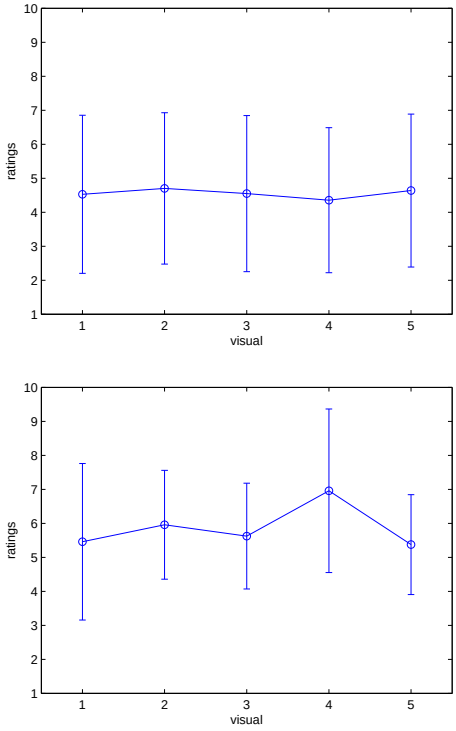


Fig. 14: Plots of the mean and ± 1 standard deviation for the visual stimuli of the audiovisual (top plot) and visual-only (bottom plot) experiments.

periment show that factor *visual* is statistically significant. This indicates that when presented in isolation, the differences between the visual stimuli are perceived more clearly and judged to be substantial but become obscure in the presence of audio stimuli. A plausible explanation is that in the context of this experiment with the products under test being loudspeakers, audio has more weight and is a more decisive factor for the product's overall performance.

The audiovisual experiment presents evidence that the influence of audio over visual is overwhelming. Unimodal experiments might have been misleading and suggest that the influence of the visual stimuli would be important for the overall evaluation.

The results in this paper are comparable to those in the benchmark experiment [2]. The ANOVA of the audiovisual experiment in this study and that of the benchmark study show that the factors have comparable effect to the model, but there is a difference in the *audio*visual* interaction which in the baseline experiment is statistically significant while in this experiment the same term is statistically non-significant. Therefore the results presented here and in the benchmark experiment suggest that audiovisual experiments might be moved to simpler setups, but the risk exists, that subtle modal interactions might be lost.

4. REFERENCES

[1] Kohlrausch, A. and van de Par, S., "Audio-visual interaction in the context of multimedia applications", in J. Blauert (Ed.), Communication Acoustics, Springer, Berlin, Germany, 2005, pp.109-138.

[2] Karandreas, A., Christensen, F., "Influence of visual appearance on loudspeaker sound quality

evaluation". Submitted to the Journal of the Audio Engineering Society, 2010.

[3] EBU Sound Quality Assessment Material CD, European Broadcasting Union, Geneva, Switzerland, 2008, Track 70: Eddie Rabbitt - "Early in the morning", timing [min.]: 0:00 - 0:09.

[4] ITU-R Rec. BS.1286, "Methods for the subjective assessment of audio systems with accompanying picture", International Telecommunications Union, Geneva, Switzerland, 1997.

[5] ITU-T Rec. P.910, "Subjective Video Quality Assessment Methods for Multimedia Applications". International Telecommunications Union, Geneva, Switzerland, 2008.

[6] ITU-R Rec. BT.500-12, "Methodology for the Subjective Assessment of the Quality of Television Pictures", International Telecommunications Union, Geneva, Switzerland, 2009.

[7] ITU-T Rec. P.911, "Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems", International Telecommunications Union, Geneva, Switzerland, 1998.

[8] ITU-T Rec. J.140, "Subjective picture quality assessment for digital cable television systems", International Telecommunications Union, Geneva, Switzerland, 1998.

[9] ITU-R Rec. BS.1116-1, "Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems", International Telecommunications Union, Geneva, Switzerland, 1997.

[10] Montgomery, D., "Design and Analysis of Experiments", 5^{th}ed., Wiley, 2001, page 234.

Manuscript D

The influence of the experimental question in audiovisual experiments.

Alex Karandreas[1], Flemming Christensen[1]

[1] *Department of Electronic Systems, Acoustics, Aalborg University, DK-9220 Aalborg, Denmark*

Correspondence should be addressed to Alex Karandreas (aka@es.aau.dk)

ABSTRACT

Audio stimuli presented through headphones were combined with photographs of loudspeakers in two subjective experiments. The only difference between experiments was the experimental question that was either neutral or referring to audio quality. A comparison between the experiments shows a small but statistically significant difference attributed to the experimental question. Results also show that the auditory modality dominates the audiovisual evaluation.

1. INTRODUCTION

A user's overall perception of a product can change when input from more than a single modality is presented, due to attention issues [1], [2] and interaction effects. However, the current literature on audiovisual experiments is scattered across many disciplines [3] and there are very few investigations on the subjective evaluations of products with audiovisual properties. To this end, the authors have previously designed a benchmark audiovisual experiment [4] in an attempt to create a useful procedure for evaluations of the overall quality of audiovisual products. The benchmark experiment featured actual loudspeakers as the source of both audio and visual stimuli.

An interesting question to raise is whether shifting from the actual loudspeaker (3D) visual presentation to a small-scale 2D presentation (a photograph of the same product shown on a small PC screen), and from the actual loudspeaker reproduction to headphone reproduction, will affect results. This is not just an interesting academic question, but could be important concerning practical applications.

The experiments described here use a similar procedure to that of the benchmark experiment, however the presentation of both audio and visual stimuli is very different. The aim of this study is to investigate the relative importance of the audio and visual stimuli as well as the validity of an alternative audiovisual stimuli presentation. An additional issue investigated in this paper is whether and to what extend the experimental question can focus the subjects attention towards one of the presented modalities. Thus 2 almost identical experiments are presented in this paper, the only difference between them being the experimental question.

2. METHOD

2.1. Experimental design

The 2 experiments presented here share a common design. Their only difference lies in the experimental question. Thus, the design will be described here once and unless explicitly stated the description is valid for both experiments.

The experiments featured the following parts in order of presentation: audition and vision screening, familiarization, audiovisual part (simultaneous presentation of audio and visual stimuli), audio-only part (no visual stimuli) and visual-only part (no audio stimuli). The presentation of the two latter parts was counter-balanced. The audio-only and visual-only parts served as baseline unimodal experiments that were compared to the bimodal audiovisual experiment.

For the audiovisual part the experimental design was a full factorial design with absolute categorical scaling. Hence, all audio stimuli were combined with all visual stimuli and each combination was presented 4 times to each subject. The order of presentations

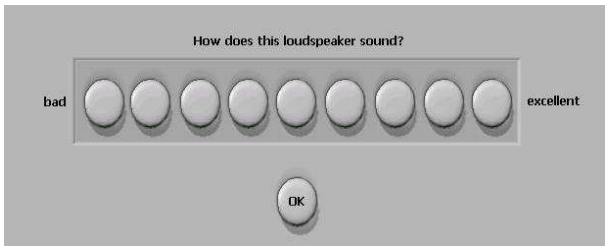


Fig. 1: Screenshot of the question and rating scale as presented in the 1^{st} experiment.

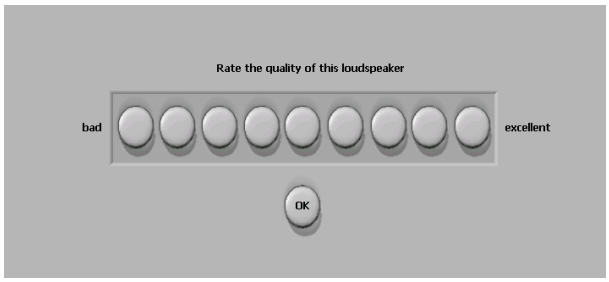


Fig. 2: Screenshot of the question and rating scale as presented in the experiment for the 2^{nd} (alternative question) experiment.

was randomized and counter-balanced across subjects. For the audiovisual part of the 1^{st} experiment, the experimental question was: *"How does this loudspeaker sound?"* and the anchors were *"bad"* and *"excellent"*. A discrete 9 point rating scale was used (figure 1). The rating scale was developed from recommendations in [5] and [6], although a 9 point scale was preferred over 5 or 7 points for higher discriminative power. The task was displayed above the rating scale as shown in figure 1.

The rationale for choosing this question was to allow the subjects to evaluate the stimuli as a whole entity. The reason for not using a question on "*audio quality*" (the question on audio quality is suggested for audio and audiovisual experiments in ITU recommendations) was to guide the subject's attention to the loudspeaker itself, so that the sound was not an isolated phenomenon but an integral part of the loudspeaker. The experiments presented here aim to evaluate the overall audiovisual impression of a product; the loudspeaker picture and the sound should be perceived as a complete product.

However, since the question contains the word *sound*, it is possible that it will direct attention towards the auditory modality. That is exactly the reason why the second experiment was performed.

For the audio-only ratings the experimental question was kept identical. For the visual-only ratings the question was: *"How would this loudspeaker sound?"*. This question is as similar to that of the audiovisual part as it could be, bearing in mind that there was no sound during this part. For the audio-only and visual-only experiments each subject was presented 4 times with all stimuli. The presentations in each experiment were randomized and counter-balanced and the order of the 2 experiments was also counter-balanced.

Note that during the audio-only and visual-only parts subjects were allowed (but not instructed) to give answers influenced by the previous audiovisual part (since they might have had associated an audio stimuli with a visual stimuli).

The experimental question for the audiovisual part of the 2^{nd} experiment (alternative question) was: *"Rate the quality of this loudspeaker"* and the anchors were again *"bad"* and *"excellent"*. The reason for choosing this question was to allow the subject to evaluate the stimuli as a whole entity without targeting their focus to specific aspects. Quality is a descriptor that can be used for audio, visual and bimodal perception studies and is recommended by ITU standards [5] and [6].

The rating scale was otherwise the same as in the 1^{st} experiment. A screenshot of the question and rating scale is shown in figure 2.

For the audiovisual, audio-only and visual-only parts of the 2^{nd} experiment the question was always the same: *"Rate the quality of this loudspeaker"*.

2.2. Audio stimuli

Both experiments featured 6 degraded versions of a single music excerpt. The excerpt features the chorus of a rock/country recording with male vocals, strumming acoustic guitar, snare drum, bass and handclaps. The music excerpt was selected from a reference recording [7] and transferred to a computer (44.1kHz, 16bit). The excerpt was carefully

Fig. 3: Pictures of loudspeakers , from left to right the visual stimuli are 1, 2, 3, 4 and 5

selected to include a complete musical phrase lasting 9 sec. Excerpts were presented at a comfortable listening level (held constant throughout the experiments and for all subjects) through circumaural headphones (BeyerDynamic DT990), which according to the manufacturer are diffused field equalized[1]. The headphones were calibrated to produce a 75dB ± 1.3dB SPL when reproducing 1/3 octave band-limited pink noise with center frequency either at 400 or 1000 Hz (-6dBFS at 44100 Hz, 16 bit), measured with a head and torso simulator (B& K HATS 4100, with B& K 4190 microphones and B& K 2669 preamplifiers, with an overall frequency range 6Hz - 20kHz). There were 3 high-pass filtered versions and 3 harmonically distorted versions. The high-pass filtered versions were filtered at 110, 220 and 440 Hz while the harmonically distorted versions were all high-pass filtered at 110 Hz and had added harmonic distortion at 3 distinct levels. The pattern of harmonic distortion was constant and the only difference was the relative level of the harmonic distortion to the 110 Hz high-pass filtered excerpt. The excerpts were not equalized with respect to loudness. The same stimuli were used in previous experiments, and were shown to be consistently ranked by a similar group of subjects[4].

2.3. Visual stimuli

In a benchmark experiment [4], 5 different loudspeaker models were selected to be the visual stimuli and the actual loudspeakers were presented. In this study instead of the actual loudspeakers, photographs were presented on a 12 inch touch screen monitor. The size of the photographs displayed on the screen was 23 x 16.5 cm. The photographs were taken in a controlled manner and the scale ratio to the actual loudspeaker was constant for all loudspeakers (see figure 3).

The loudspeakers used in this study were:

- Satellite (of a surround system) 1-way unit in grey plastic cabinet. Dimensions: 12.5 x 9 cm. Diaphragm not visible.
- Large bookshelf 3-way loudspeaker with 4:3 cabinet proportions. Dimensions: 29 x 41 cm. Diaphragm not visible.
- Large bookshelf 2-way unit with a rectangular wooden cabinet. Dimensions: 35 x 23 cm. Diaphragm visible.
- Floor standing 4-way loudspeaker. Dimensions: 184 x 18.5 cm. Diaphragm not visible.
- Small bookshelf 1-way unit in black plastic cabinet with a tilted upper section. Dimensions: 20.5 x 13 cm. Diaphragm not visible.

.

2.4. Screening

6 naive[2] university students participated in the 1^{st} experiment (3 male and 3 female, mean age = 21.3

[1]ITU recommendations [8] and [9] recommend diffuse field equalization for headphones for subjective evaluations.

[2]In the context of this paper, a naive subject is one that has no prior experience from viewing or listening tests and has limited knowledge of loudspeakers (technical and commercial).

yrs, std = 2.25) and another group of 6 naive university students in the 2^{nd} experiment (3 male and 3 female, mean age = 24.8 yrs, std = 3.5) An audiometric screening was made, and none had a hearing loss greater than 15dB HL in either ear at any octave band frequency between 125 Hz and 8 kHz. The participants were required to have normal or corrected to normal vision acuity and normal color vision. The participants vision was inspected prior to the experiments using standard vision charts. Concerning acuity, no error on the 20/30 line of the standard eye chart was made. Concerning color vision, no plates were missed out of 12 on an Ishihara test (with 2 out of 12 misses being the criteria) [5], [6], [10], [11].

As a further supplement, prior to the experiment, data regarding the listening habits and prior experience of the participants were collected by means of a questionnaire to ensure that the subject were naive listeners.

2.5. Familiarization

A familiarization procedure introduced subjects to the stimuli and the experimental procedure. In 2 different familiarization sessions all visual and then all audio stimuli were presented in isolation. A third session featured selected audiovisual combinations presented in the same way as in the actual experiment.

More precisely, during the first familiarization the visual stimuli were presented to the subjects without any reference to the rating scale or anchors.

The second familiarization, introduced the range of audio degradations, presenting the least and most degraded excerpts. In order to present subjects with the range of audio degradations, only during this familiarization the audio stimuli were labeled. The least degraded stimulus was termed "excellent" and the most degraded stimuli (of both degradation methods) were termed "bad".

Subjects were instructed that these stimuli were just some of the audio stimuli they would hear during the experiment.

A third familiarization emulated the stimuli presentation and the evaluation procedure as it would take place during the actual experiment. Three audiovisual presentations were given, during which the least degraded stimulus, and the two most degraded

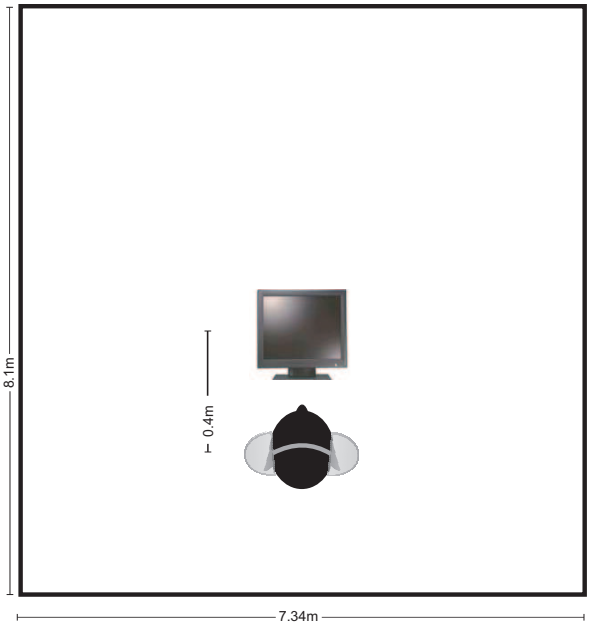


Fig. 4: Setup in the listening room.

stimuli (high-pass filtering for one excerpt, harmonic distortion for the other) were coupled with a loudspeaker photograph. The presented loudspeaker was also counter-balanced among subjects.

2.6. Setup

The experiment was conducted in a laboratory room conforming with the ITU-R BS.1116 recommendation for listening rooms suitable for evaluations of multichannel audio systems [9], and controlled from an adjacent control room. The setup can be seen in figures 4 and 5.

The visual stimuli were shown on a touch screen (ELO Touchsystems ETL12IC, diagonal size: 12 inches) in front of the subjects. The size of the photographs displayed on the screen was 23 x 16.5 cm. The screen was adjusted on a stand at a height of 70 cm from the floor, an angle of 30^0 with respect to the floor and a viewing distance of approximately 40 cm.

The experiment was fully automated and controlled by a PC running custom-made software (programmed in Labview 6.1). This included the order of presentation of the stimuli (randomization), the stimuli generation and the data collection. The

Table 1: Light and viewing conditions in the laboratory room. Luminance measured with Gossen MAVOLUX 5032C, Class C acc. DIN 5032-7, Min. Sensitivity: $0.1lx$.

Background room illumination	0 lx
Listening room dimensions	$8.1 * 7.34 * 2.86$ m
Viewing distance to the touch screen	0.4 m

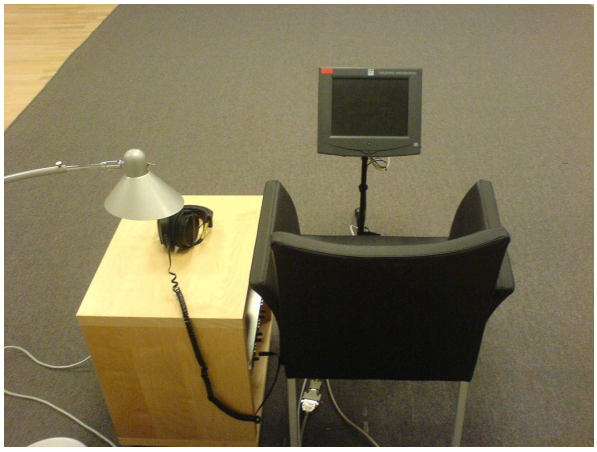


Fig. 5: Photograph of the setup. The headphones and headphone amplifier are also visible.

audio stimuli were generated by an internal sound card (RME Digi 9636, 24-bit, 96 kHz), were converted to analog (Tracer Technologies Big Daddi, 24 bit D/A converter), fed to a headphone amplifier (Behringer Powerplay Pro-XL HA4700) and reproduced by headphones (Beyerdynamic DT 990 PRO circumaural, diffuse field equalized headphones).

The listening room was kept completely dark. The light and viewing conditions in the laboratory room are shown in table 1.

3. RESULTS

This section initially presents the effect of the experimental question in the 2 experiments, followed by the individual results for each experiment as well as a general comparison across the results of the 2 experiments.

3.1. Effect of the experimental question

To examine the effect of the experimental question, the data from the 2 experiments were pooled together and the effect of the experimental question was modeled as factor *experiment* in the ANOVA analysis (see figure 6).

Factor *experiment* and the interaction *audio*experiment* are statistically significant. The interaction term indicates that there were differences in the audio evaluations among the 2 experiments that could be attributed to the experimental question.

Figures 7, 8, 9 and 10 show that although differences exist between the 2 experiments, these differences are very small. In the case of figure 8 for example, the ratings of the 2^{nd} experiment are overall decreased about half a rating point and the ranking of the visual stimuli is different, but for each of the 2 experiments the differences between the levels of factor *visual* are within 0.25 rating points. Overall, the largest difference between the 2 experiments lies in the visual-only evaluations (figure 10), a result that could be partly attributed to the different experimental questions of the visual-only evaluations.

As mentioned in the introduction, a benchmark experiment [4] was conducted prior to the experiments described here. The results of that experiment showed that the auditory modality dominated the audiovisual evaluation. The experimental question *"How does this loudspeaker sound?"* that refers directly to one of the 2 modalities might constitute a bias. Assuming that the experimental question influenced the results, it would be reasonable to expect that modifying the experimental question to be completely neutral, would result in a greater influence of the visual modality. The results presented here show that the effect of the experimental question is statistically significant, however the impact of visual appearance on the results has not changed

Analysis of Variance

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
A	4519.03	5	903.805	307.8	0
V	3.15	4	0.788	0.27	0.8983
Exp	13.48	1	13.481	4.59	0.0323
A*V	32.91	20	1.646	0.56	0.9398
A*Exp	102.47	5	20.495	6.98	0
V*Exp	7.48	4	1.871	0.64	0.6361
A*V*Exp	41.3	20	2.065	0.7	0.8261
Error	4052.09	1380	2.936		
Total	8777.1	1439			

Constrained (Type III) sums of squares.

Fig. 6: ANOVA table for the pooled audiovisual data of both experiments, including 2 and 3-way interactions. *A*, *V* and *Exp* refer to factors *audio*, *visual* and *experiment* respectively.

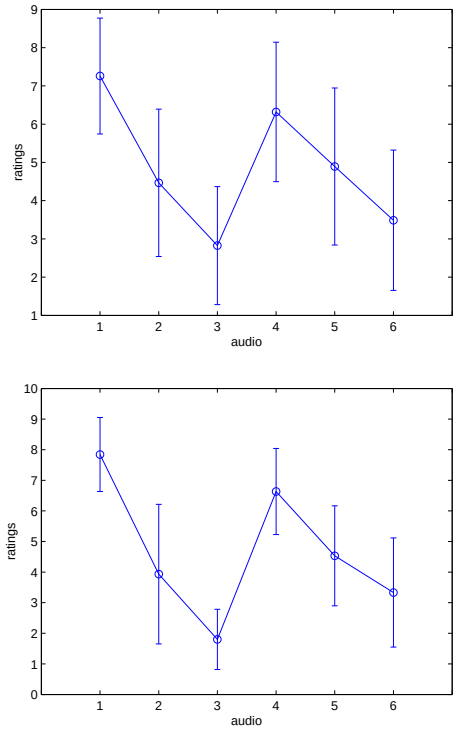


Fig. 7: The averaged results for the audiovisual experiment with respect to the audio stimuli. Top plot shows means ± standard deviations of the data of the 1^{st} experiment. Bottom plot shows data of the 2^{nd} experiment. A rating equal to 1 corresponds to low quality and 9 to high quality. Audio stimuli 1, 2 and 3 are high-pass filtered stimuli and stimuli 4, 5 and 6 are harmonically distorted stimuli. Note that the y-axis are different.

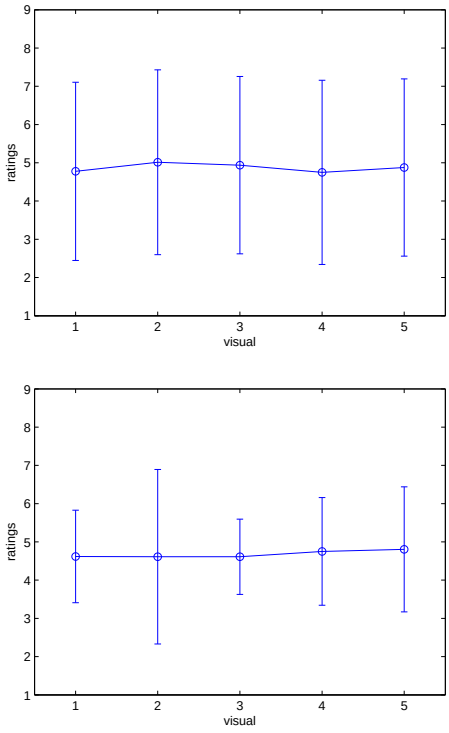


Fig. 8: The averaged results for the audiovisual experiment with respect to the visual stimuli. Top plot shows means ± standard deviations of the data of the 1^{st} experiment. Bottom plot shows data of the 2^{nd} experiment.

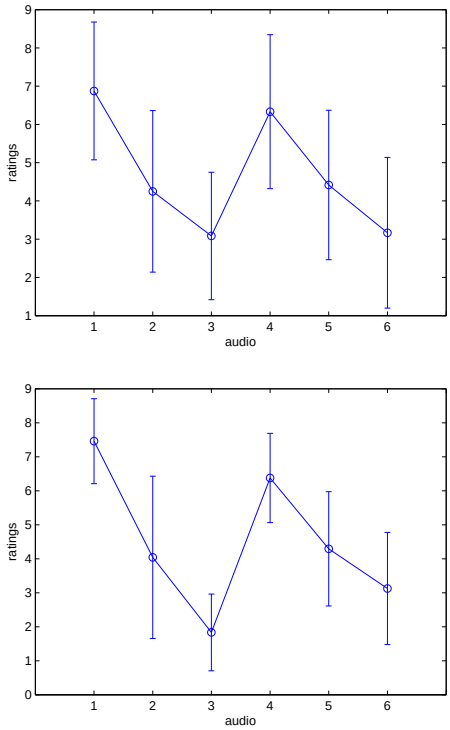


Fig. 9: Audio-only data. Top plot shows data from the 1^{st} experiment. Bottom plot shows data from the 2^{nd} experiment.

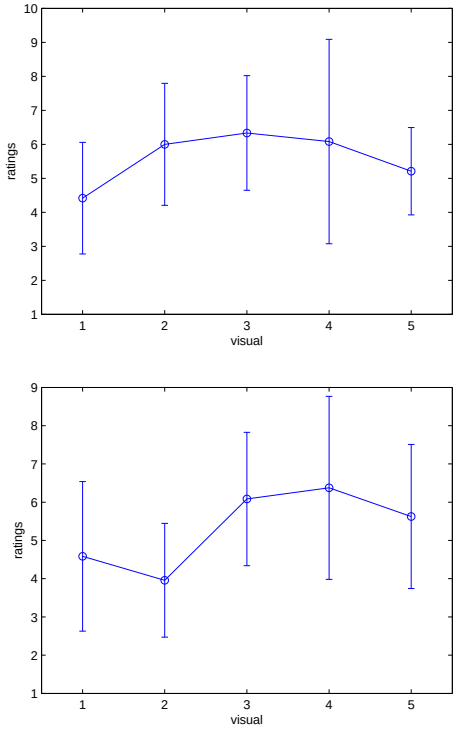


Fig. 10: Visual-only data. Top plot shows data from the 1^{st} experiment. Bottom plot shows data from the 2^{nd} experiment. Note that the y-axis are different.

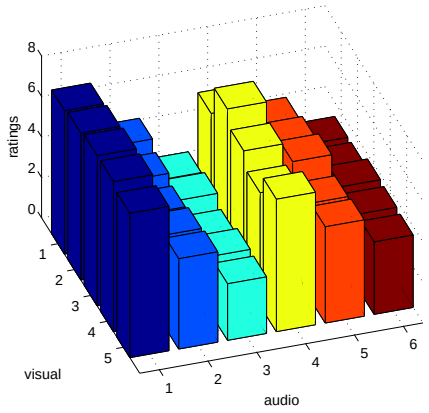


Fig. 11: Average ratings across subjects for all audiovisual presentations in the 1^{st} experiment. A rating equal to 1 corresponds to low quality and 9 to high quality. Audio stimuli 1, 2 and 3 are high-pass filtered stimuli and stimuli 4, 5 and 6 are harmonically distorted stimuli.

(the *visual*experiment* interaction is not statistically significant and the Sum of Squares and P value of factor *visual* in both experiments presented in this paper is very similar), and the differences between the results of the 2 audiovisual experiments are very small. These results show that the influence of neither the audio or visual modalities has changed considerably.

3.2. Results for the 1^{st} experiment

3.2.1. Audiovisual part results, 1^{st} experiment

The average results across subjects for all audiovisual presentations are shown in figure 11. Large differences are seen on the ratings along the audio axis whereas the ratings along the visual axis show only small differences, that are substantial only for audio level 4 (*a4*).

The ANOVA table is shown in figure 12. Factors *subjects*, *audio* and the 2-way interaction *subjects*audio* are statistically significant. The ANOVA analysis suggests that factor *visual* has negligible influence to the results. The ratio of the Sum of Squares of each factor to the total Sum of Squares can be expressed as a percentage contribution of each factor to the ANOVA model [12]. For

Analysis of Variance

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
S	465.44	5	93.089	47.85	0
A	1677.42	5	335.485	172.46	0
V	6.18	4	1.544	0.79	0.5294
S*A	566.46	25	22.658	11.65	0
S*V	33.42	20	1.671	0.86	0.6405
A*V	43.08	20	2.154	1.11	0.3372
S*A*V	138.42	100	1.384	0.71	0.9816
Error	1050.47	540	1.945		
Total	3982.99	719			

Constrained (Type III) sums of squares.

Fig. 12: The ANOVA table of the audiovisual data in the 1^{st} experiment including 2 and 3-way interactions. *S*, *A* and *V* refer to factors *subjects*, *audio* and *visual* respectively.

this experiment, factor *audio* accounts for 42% of the variability of the experiment, while factor *visual* accounts for 0,15% and the interaction term *subject*audio* accounts for 14%.

The *subject*audio* interaction (figure 13) shows that there are large differences in the ratings between the audio levels. The figure also shows that there is between-subjects variance, however the same ranking is followed by all subjects in all but 2 cases: the interaction between audio level 3 and subject 6 (*a3*s6*), where *a3* is rated lower than *a2* and *a4*s1* where *a4* is rated lower than *a2*. The *subject*audio* interaction is also shown in figure 14. The *subject*audio* interaction shows that the audio levels have a strong influence on the results (the plot follows the pattern of the audio stimuli where *a1* and *a4* are the least degraded versions of either degradation methods while *a2*, *a3*, *a5* and *a6* are progressively more degraded versions of either degradation method). The influence of factor *subjects* is also large.

The *subject*visual* interaction (figure 15) shows that factor *subjects* is more influential than factor *visual* which has only a minimal effect.

Subject differences are to be expected since no reference and minimal training is used in the experiment.

Statistical analysis of the data showed that the data are not normally distributed but exhibit less variance than expected. Some evidence indicates that this could be due to the characteristics of the rating scale. Attempts to normalize the data had minimal

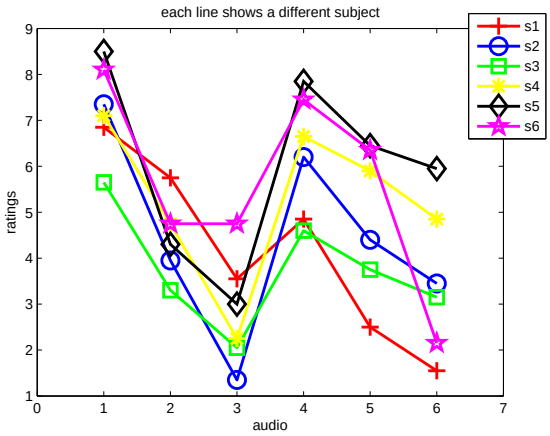


Fig. 13: Plot of the 2-way interaction between factors *subject* and *audio*, for the audiovisual data. 1^{st} experiment.

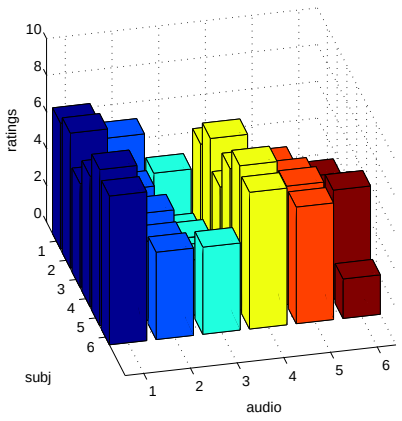


Fig. 14: Plot of the 2-way interaction between factors *subject* and *audio*. Audiovisual data. 1^{st} experiment.

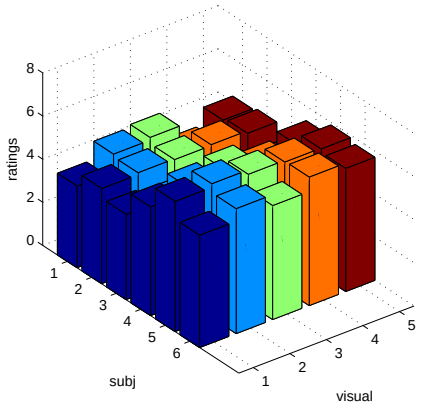


Fig. 15: Plot of the 2-way interaction between factors *subject* and *visual*. Audiovisual data. 1^{st} experiment.

effect and did not change the shape of the distribution. Non-parametric analysis on the raw data showed results to be very similar to those of the ANOVA (the same factors are statistically significant).

3.2.2. Audio part results, 1^{st} experiment

The average results across subjects for all audio-only presentations are shown in figure 16. Large differences are seen on the ratings along the audio axis, however the differences between the 2 degradation methods are small. The maximum difference between levels is about 4 rating points, similar to the maximum difference for the same factor in the audiovisual experiment.

The ANOVA table is shown in figure 17. Factors *subjects*, *audio* and the 2-way interaction are all strongly significant.

3.2.3. Visual part results, 1^{st} experiment

The average results across subjects for all visual-only presentations are shown in figure 18. The maximum difference between levels is about 2 rating points. 3 visual levels (*v2*, *v3* and *v4*) are rated alike.

The ANOVA table is shown in figure 19. Both factors visual and subjects and the 2-way interaction

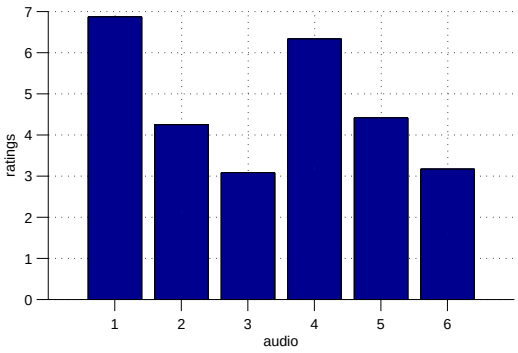


Fig. 16: Average ratings across subjects for all audio-only presentations. A rating equal to 1 corresponds to low quality and 9 to high quality.

Analysis of Variance

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
S	106.646	5	21.3292	11.03	0
A	303.479	5	60.6958	31.4	0
S*A	196.062	25	7.8425	4.06	0
Error	208.75	108	1.9329		
Total	814.938	143			

Constrained (Type III) sums of squares.

Fig. 17: The ANOVA table for the audio-only part of the 1^{st} experiment.

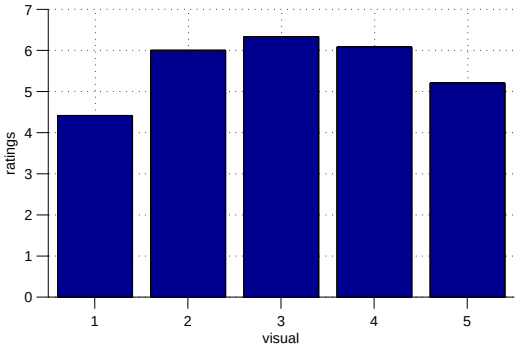


Fig. 18: Average ratings across subjects for all visual-only presentations. A rating equal to 1 corresponds to low quality and 9 to high quality.

Analysis of Variance

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
S	69.242	5	13.8483	26.95	0
V	59.633	4	14.9083	29.01	0
S*V	331.467	20	16.5733	32.25	0
Error	46.25	90	0.5139		
Total	506.592	119			

Constrained (Type III) sums of squares.

Fig. 19: The ANOVA table for the visual-only part of the 1^{st} experiment.

are statistically significant with the Sum of Squares for the 2-way interaction being much larger than that of the main terms.

3.2.4. Data across parts for the 1^{st} experiment

The overall ranking of audio and visual stimuli for the audiovisual as well as the audio-only and visual-only parts of the 1^{st} experiment are shown in table 2. The audio stimuli are identically ranked in the audiovisual and audio-only experiments, while the visual stimuli ranking between experiments is different. Figures 20 and 21 show the means $\pm$ standard deviations of the data in the 3 experiments. Close resemblance is seen for the audio ratings in the audiovisual and audio-only experiments, while visual ratings are different between the audiovisual and visual-only experiments. A difference between levels with visual stimuli 1 and 5 having a lower rating than the rest can be seen for the visual-only experiment. In the audiovisual experiment there are hardly any differences between visual levels.

3.3. Results for the 2^{nd} experiment

3.3.1. Audiovisual part results, 2^{nd} experiment

The average results across subjects for all audiovisual presentations are shown in figure 22. Large differences are seen on the ratings along the audio axis whereas the ratings along the visual axis show only small differences. For factor *audio*, there is a maximum difference of about 6 rating points with audio stimuli 3 being rated lowest and audio stimuli 1 being rated highest. For factor *visual* the maximum difference is about 1 rating point.

The ANOVA analysis (figure 23) suggests that factors *subjects*, *audio* and the 2-way interaction *sub-*

Table 2: Data across parts, averaged across subjects for the 1^{st} experiment. The ranking order is shown from lowest to highest. A(AV) and V(AV) are the averaged results for the audiovisual part with respect to the audio and visual stimuli respectively.

experiment	A(AV)	V(AV)	Audio-only	Visual-only
ranking	3,6,2,5,4,1	1,4,5,3,2	3,6,2,5,4,1	1,5,2,4,3

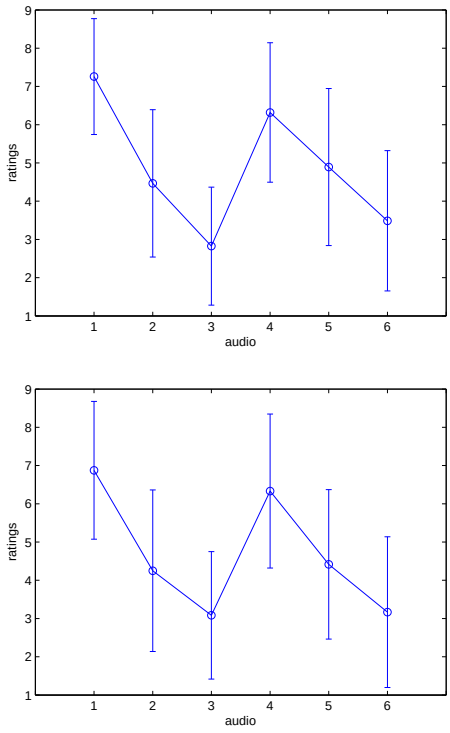


Fig. 20: Plots of the mean and ± 1 standard deviation for the audio stimuli of the audiovisual (top plot) and audio-only (bottom plot) parts. 1^{st} experiment.

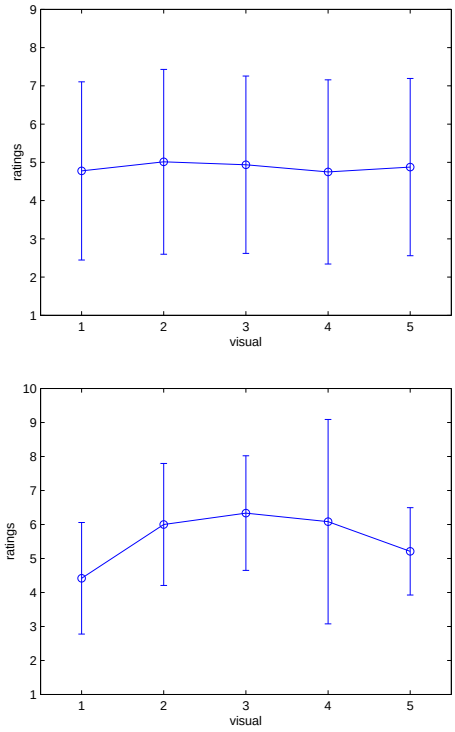


Fig. 21: Plots of the mean and ± 1 standard deviation for the audio stimuli of the audiovisual (top plot) and visual-only (bottom plot) parts. 1^{st} experiment.

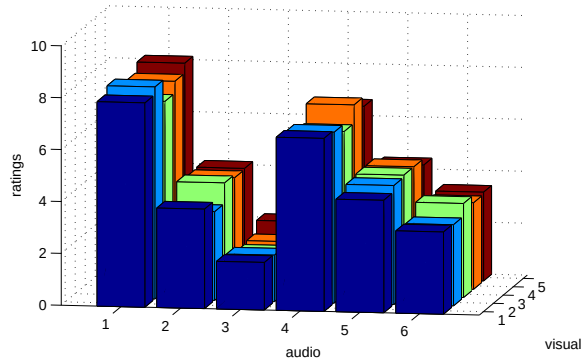


Fig. 22: Average ratings across subjects for all audiovisual presentations of the 2^{nd} experiment. A rating equal to 1 corresponds to low quality and 9 to high quality.

Analysis of Variance

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
S	279.85	5	55.969	34.37	0
A	2939.83	5	587.966	361.1	0
V	4.89	4	1.224	0.75	0.5573
S*A	530.66	25	21.227	13.04	0
S*V	25.52	20	1.276	0.78	0.7346
A*V	33.29	20	1.664	1.02	0.4332
S*A*V	87.59	100	0.876	0.54	0.9999
Error	879.25	540	1.628		
Total	4780.89	719			

Constrained (Type III) sums of squares.

Fig. 23: The ANOVA table for the audiovisual data of the 2^{nd} experiment including 2 and 3-way interactions.

*jects*audio* are statistically significant while the influence of factor *visual* is not significant. Furthermore, factor *audio* seems to be by far the most influential factor. From the ANOVA table it can be seen that factor *audio* accounts for 61% of the variability of the experiment, while factor *visual* accounts for 0,1%, factor *subjects* accounts for 5% and the interaction term *subject∗audio* accounts for 11%.

The *subject*audio* interaction (figures 24 and 25) shows that there are between-subjects differences, however the same trend is followed by all subjects and there are large differences in the ratings between the audio levels.

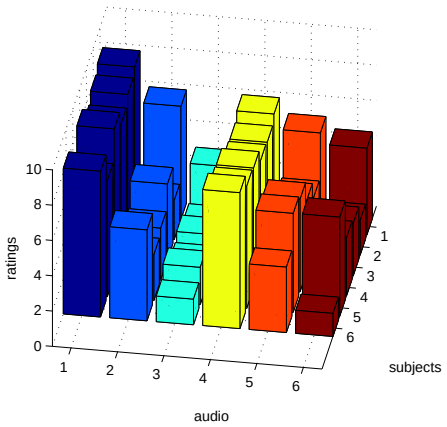


Fig. 24: Plot showing the interaction between factors *audio* and *subjects*. Audiovisual data. 2^{nd} experiment.

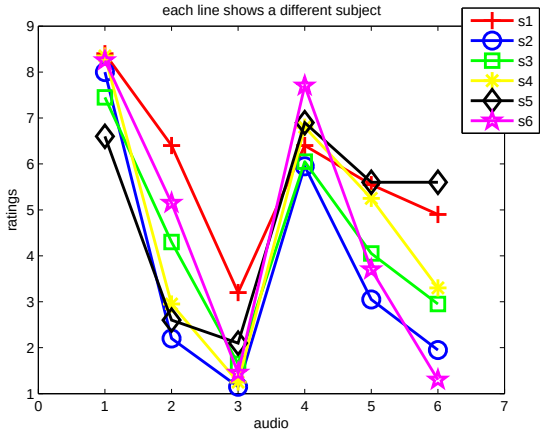


Fig. 25: Plot of the 2-way interaction between factors *subject* and *audio*, for the audiovisual data. 2^{nd} experiment.

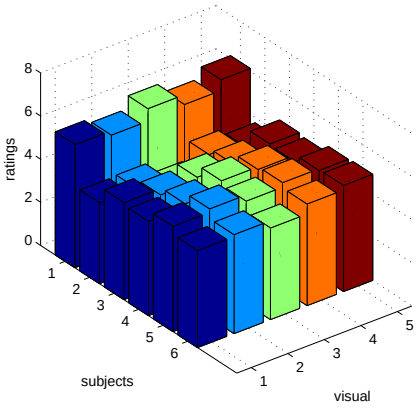


Fig. 26: Plot showing the interaction between factors *visual* and *subjects*. Audiovisual data. 2^{nd} experiment.

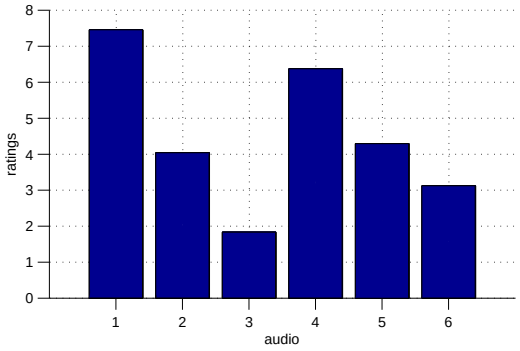


Fig. 27: Average ratings across subjects for all audio-only presentations. A rating equal to 1 corresponds to low quality and 9 to high quality.

The *subject*visual* (figure 26) interaction responses for subjects 3, 4, 5 and 6 are similar, while subject 1 and 2 give overall more elevated or decreased responses respectively. Factor *visual* shows little influence.

Similarly to the previous experiment, statistical analysis showed that the data are not normally distributed and attempts to normalize the data had little effect. Non-parametric analysis on the raw data (same data used for the ANOVA) showed results to be very similar to those of the ANOVA.

3.3.2. Audio part results, 2^{nd} experiment

The average results across subjects for all audio-only presentations are shown in figure 27. Large differences are seen on the ratings along the audio axis.

Factors *subjects*, *audio* and the 2-way interaction are all strongly significant (figure 28). From the ANOVA table we see that factor *audio* accounts for 58% of the variability of the experiment, while factor *subject* accounts for 4% and the interaction term *subject*audio* accounts for 21%. Thus most of the variance is accounted for by factor *audio* and the influence of factor *subject* is relatively much smaller.

Analysis of Variance

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
S	41.479	5	8.296	6.53	0
A	516.479	5	103.296	81.28	0
S*A	184.729	25	7.389	5.81	0
Error	137.25	108	1.271		
Total	879.938	143			

Constrained (Type III) sums of squares.

Fig. 28: The ANOVA table for the audio-only part of the 2^{nd} experiment.

3.3.3. Visual part results, 2^{nd} experiment

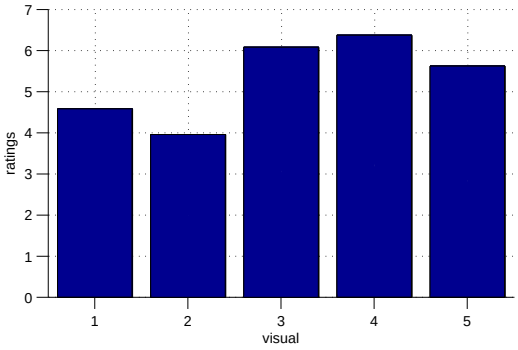


Fig. 29: Average ratings across subjects for all visual-only presentations. A rating equal to 1 corresponds to low quality and 9 to high quality.

Analysis of Variance

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
S	37.875	5	7.575	18.55	0
V	100.45	4	25.1125	61.5	0
S*V	347.25	20	17.3625	42.52	0
Error	36.75	90	0.4083		
Total	522.325	119			

Constrained (Type III) sums of squares.

Fig. 30: The ANOVA table for the visual-only part of the 2^{nd} experiment.

The average results across subjects for all visual-only presentations are shown in figure 29. The figure shows that 3 levels are rated closely while levels 1 and 2 are rated lower, with a maximum difference of about 2 rating points.

The ANOVA table is shown in figure 30. Factors *subjects*, *visual* and the 2-way interaction are all strongly significant.

3.3.4. Data across parts for the 2^{nd} experiment

The overall ranking of audio and visual stimuli for the audiovisual as well as the audio-only and visual-only parts are shown in table 3. The audio stimuli are almost identically ranked in the audiovisual and audio-only experiments, while the visual stimuli ranking between experiments is different. Figures 31 and 32 show the means ± standard deviations of

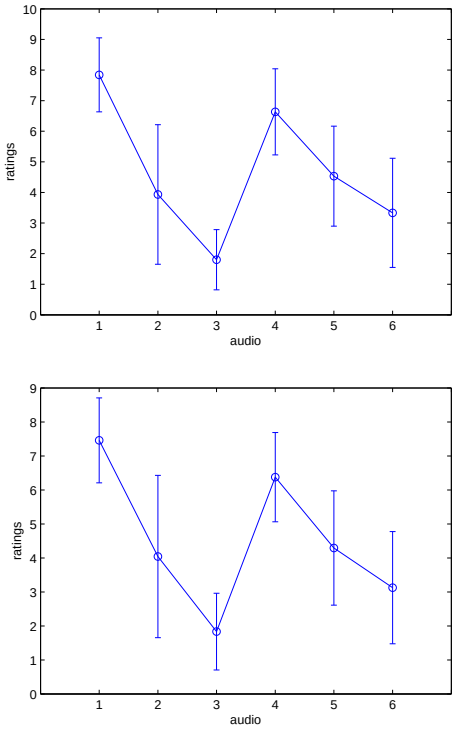


Fig. 31: Plots of the mean and ± 1 standard deviation for the audio stimuli of the audiovisual (top plot) and audio-only (bottom plot) parts. 2^{nd} experiment.

the data. Close resemblance is seen for the audio ratings, while mean visual ratings are different between the audiovisual and visual-only experiments. For the mean visual ratings in the visual-only experiment there is a difference between levels, while in the audiovisual experiment there are hardly any differences between visual levels.

4. CONCLUSIONS

4.1. Main conclusions

- The results of the 1^{st} and 2^{nd} experiments are similar. The experimental question had a statistically significant but small effect. The influence of both the audio and visual stimuli remained largely unaltered.
- All results indicate that for this study *audio* was

Table 3: Data across parts, averaged across subjects for the 2^{nd} experiment. The ranking order is shown from lowest to highest. A(AV) and V(AV) are the averaged results for the audiovisual part with respect to the audio and visual stimuli respectively.

experiment	A(AV)	V(AV)	Audio-only	Visual-only
ranking	3,6,2,5,4,1	2,3,1,4,5	3,6,2,5,4,1	2,1,5,3,4

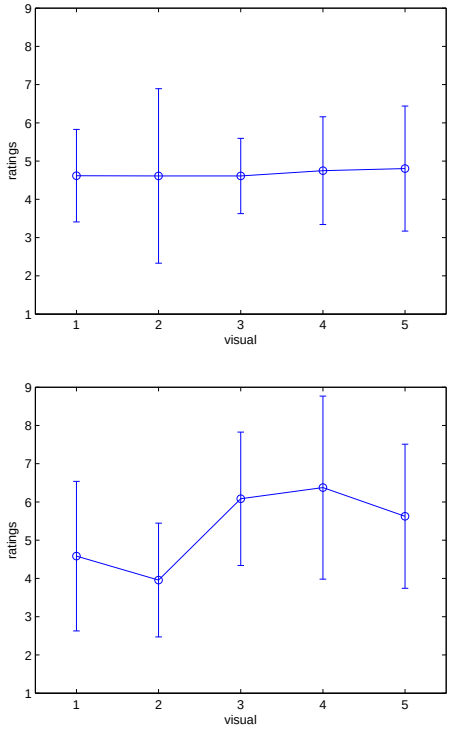


Fig. 32: Plots of the mean and ± 1 standard deviation for the audio stimuli of the audiovisual (top plot) and visual-only (bottom plot) parts. 1^{st} experiment. 2^{nd} experiment.

the primary factor while *visual* had minimal influence.

- The results concerning factors *audio* and *visual* are similar to those in similar experiments featuring actual loudspeaker reproduction.
- Overall, the results are comparable to those in similar experiments where the audio and visual presentation was from actual loudspeakers, showing that alternative reproduction techniques are valid in audiovisual experiments.
- The results of the bimodal experiments cannot be directly inferred from the unimodal experiments, suggesting that unimodal experiments can lead to misleading conclusions. This study did not show any particular interactions between the audio and visual stimuli, however it showed that under these specific experimental conditions, design and with the selected stimuli, the audio stimuli are the dominating factor in the audiovisual evaluation whereas visual stimuli have a minimal effect.

The results in this paper are comparable to those in the benchmark experiment [4]. The rankings and ratings of the audio and visual stimuli for the audiovisual presentations are similar to the benchmark experiment. The ANOVA in the benchmark experiment and the experiments presented here show that the factors have comparable effect to the model, but there is a difference in the *audio*visual* interaction which in the benchmark experiment is statistically significant with $P = 0.0009$ while in the experiments presented here the same term is statistically non-significant with $P = 0.3372$ and $P = 0.4332$ in the 1^{st} and 2^{nd} experiment respectively. Thus, under the given circumstances audiovisual experiments

might be moved to simpler setups, but the risk exists, that subtle modal interactions might be lost.

4.2. 1^{st} experiment

For factor *audio* the rank of the levels in the audiovisual and the audio only part is exactly the same. This suggests that the difference between the audio stimuli is large and easily perceived by the subjects.

For factor *visual* the order of the levels in the audiovisual and the visual only part is different (the order is 1-4-5-3-2 and 1-5-2-4-3 in the audiovisual and visual-only part respectively). Interestingly, levels 1 and 5 that are ranked low in both cases represent the 2 smallest loudspeakers used in this study. Furthermore, the ratings of the visual levels in the audiovisual part are very similar while in the visual-only part there are clear differences. These results might suggest that in isolation the differences between the visual stimuli are perceived more clearly but become obscure when combined with audio. A plausible explanation is that in this context *audio* has more weight and is a more decisive factor for the products overall performance.

4.3. 2^{nd} experiment

For the 2^{nd} experiment the rank of the audio stimuli in the audio-only part and the audiovisual part is the same. For the audio-only part, the maximum difference between ratings for factor *audio* is about 4 rating points, similar to the maximum difference for the same factor in the audiovisual part. This suggests that the difference between the audio stimuli is large and easily perceived by the subjects and that the effect of factor *visual* is very small.

For the visual-only part, for factor *visual* the maximum difference between ratings is about 2 rating points, quite larger than the difference in the audiovisual part which is less than half a rating point. The ranking of the levels for this factor in the audiovisual and the visual only part is different (the order is 2-3-1-4-5 and 2-1-5-3-4 in the audiovisual and visual-only parts respectively). Comparing to the visual-only rankings of the 1^{st} experiment (1-5-2-4-3) we see that there are differences, but loudspeakers 1 and 5 are still ranked low while the largest loudspeaker (4) is ranked highest or next to highest.

5. REFERENCES

[1] Alais, D., Morrone, C. and Burr, D., "Separate attentional resources for vision and audition", Proc. Biol. Sci., 273(1592), 1339-1345, 2006.

[2] Massaro, D.W. and Warner, D.S, "Dividing attention between auditory and visual perception", Percept. Psychophys., 21, 569-574, 1977.

[3] Kohlrausch, A. and van de Par, S., "Audiovisual interaction in the context of multimedia applications", in J. Blauert (Ed.), Communication Acoustics, Springer, Berlin, Germany, 2005, pp.109-138.

[4] Karandreas, A., Christensen, F., "Influence of visual appearance on loudspeaker sound quality evaluation". Submitted to the Journal of the Audio Engineering Society, 2010.

[5] ITU-T Rec. P.910, "Subjective Video Quality Assessment Methods for Multimedia Applications". International Telecommunications Union, Geneva, Switzerland, 2008.

[6] ITU-T Rec. P.911, "Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems", International Telecommunications Union, Geneva, Switzerland, 1998.

[7] EBU Sound Quality Assessment Material CD, European Broadcasting Union, Geneva, Switzerland, 2008, Track 70: Eddie Rabbitt - "Early in the morning", timing [min.]: 0:00 - 0:09.

[8] ITU-R Rec. BS.1284-1, "General methods for the subjective assessment of sound quality", International Telecommunications Union, Geneva, Switzerland, 2003.

[9] ITU-R Rec. BS.1116, "Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems", International Telecommunications Union, Geneva, Switzerland, 1997.

[10] ITU-R Rec. BT.500-12, "Methodology for the Subjective Assessment of the Quality of Television Pictures", International Telecommunications Union, Geneva, Switzerland, 2009.

[11] ITU-T Rec. J.140, "Subjective picture quality assessment for digital cable television systems", International Telecommunications Union, Geneva, Switzerland, 1998.

[12] Montgomery D., "Design and Analysis of Experiments", 5^{th}ed., Wiley, 2001, page 234.

Manuscript E

Subjective audiovisual assessment of loudspeakers.

Alex Karandreas[1]

[1] *Department of Electronic Systems, Acoustics, Aalborg University, DK-9220 Aalborg, Denmark*

Correspondence should be addressed to Alex Karandreas (aka@es.aau.dk)

ABSTRACT

Multisensory experiments lack a common methodology which would enable comparisons between studies. The author encountered a number of issues in previous audiovisual evaluations. Some of these issues were the complexity of experimental design, stimuli selection and realism for bimodal presentations. This study investigates a different design approach that aims to simplify these issues. Loudspeaker photographs combined with a range of audio excerpts are presented to subjects who are asked to rate the overall quality of the audiovisual presentation. Audio-only and visual-only evaluations are also collected for the same stimuli and compared to the audiovisual evaluations. Results show that both the audio and visual stimuli significantly influence the overall audiovisual evaluations.

1. INTRODUCTION

Subjective evaluation experiments of products usually consider the input from a single modality. However, for most products the overall perception depends on more than one modality. The present study evaluates overall impression in relation to audition and vision, using loudspeakers as an example. In order to quantify the bias that the loudspeaker appearance has on sound quality evaluation, music excerpts are coupled with loudspeaker photographs.

In an attempt to arrive to a useful methodology for audiovisual experiments, the author has previously designed a series of audiovisual experiments that shared a common design and investigated stimuli selection and stimuli presentation techniques as well as the choice of experimental question [1].

The author's previous audiovisual experiments featured audio stimuli that were degraded while the visual stimuli were not. This might have caused an imbalance between the 2 modalities [2] and subjects might have payed more attention towards the audio stimuli in their effort to distinguish between the different degradation levels rather than focusing equally to both audio and visual stimuli.

Another issue concerning the experimental design of the previous experiments was that all audio stimuli were combined with all visual stimuli (in a full factorial design), which could have led subjects to believe that the audiovisual combinations were not representative for each pair; in other words that the choice of audiovisual pairs was random. This issue was possibly intensified by the familiarization sessions before the actual experiment. During 2 different familiarization sessions all visual and then all audio stimuli were presented in isolation. It is therefore possible that this could again have led subjects to think that the audio and visual stimuli were 2 isolated phenomena.

These issues are crucial because the aim of those studies was to evaluate the overall audiovisual impression of a product; the loudspeaker photograph and the audio were to be thought of as a unique product.

This paper introduces the reasoning behind a different experimental design, describes the experiment that was conducted and presents the experimental results. Finally, a discussion of the results in contrast with results obtained in previous experiments is made together with a discussion of the validity of the chosen experimental design and its usefulness for further audiovisual investigations.

2. METHOD

2.1. Experimental design

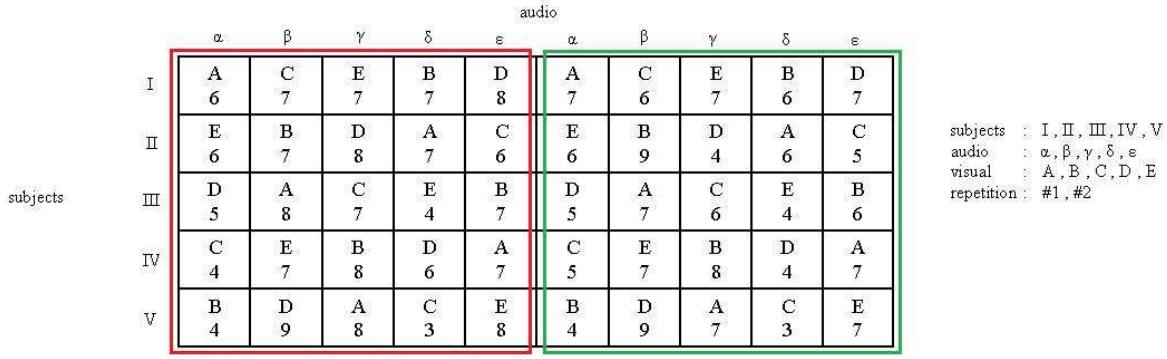


Fig. 1: The 2 Latin Squares (LS) presented in the experiment. The 2 LS are highlighted by red and green squares. Results are also shown. Note that the second repetition is re-arranged here for clarity - for the experiment the presentations of the repetition (the LS on the right) were randomized. The legend shows the levels of each factor.

For this experiment instead of combining all audio stimuli with all visual stimuli, each subject was presented with a unique set of stimuli where each audio stimulus was combined exclusively with one and only one visual stimulus. This was done with the use of Latin Squares (LS).

A LS of order 5 was chosen in order to allow for a 5x5 arrangement of 5 audio stimuli and 5 visual stimuli resulting in 25 audiovisual combinations. Each subject was presented only with 1 row of the LS. Since the audiovisual combinations are unique for each subject, that means that 5 subjects are required to collect 1 data point (rating) per audiovisual combination. In order to increase statistical power it is usual to have at least one repetition per audiovisual combination (2 data points per audiovisual combination). Thus the same set of audiovisual combinations were repeated for each subject, and the 2^{nd} time the set was randomized and care was taken never to have 2 consecutive identical stimuli. The LS used in this experiment is shown in figure 1.

This is a completely additive model. The repetition is done keeping the row and column variables and levels identical (for more information see [3] page 149).

For the audio-only and visual-only experiments that followed the audiovisual experiment, the stimuli presentation for each subject was randomized and counter-balanced and there were 2 presentations of each stimulus [1].

The current experiment included the following parts in order of presentation: audition and vision screening, familiarization, audiovisual experiment (simultaneous presentation of audio and visual stimuli), audio-only experiment (no visual stimuli) and visual-only experiment (no audio stimuli). The presentation of the two latter experiments was counterbalanced across subjects.

5 university students (3 male and 2 female) participated in this study (mean age = 24.2 yrs, std = 2.3). An audiometric screening was made, and none

[1] To ensure a randomized and counter-balanced presentation the 5 stimuli were arranged as 2 5x5 LS and each subject was presented with 1 row from each LS. Care was taken not to have repeating trials.

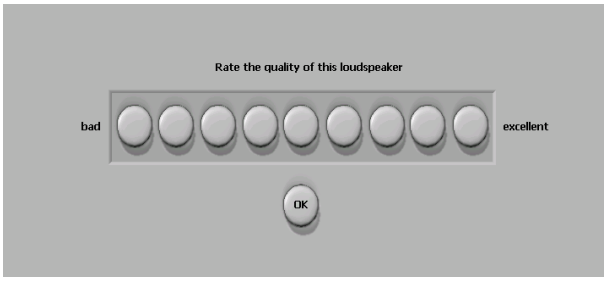


Fig. 2: Screenshot of the question and rating scale as presented in the experiment.

had a hearing loss greater than 15dB HL in either ear at any octave band frequency between 125 Hz and 8 kHz. The participants were required to have normal or corrected to normal vision acuity and normal color vision in accordance to ITU recommendations [4], [5], [6], [7]. The participants vision was inspected prior to the experiments using standard vision charts. Concerning acuity, no error on the 20/30 line of the standard eye chart was made. Concerning color vision, no more than 2 plates were missed out of 12 on an Ishihara test.

As a further supplement, prior to the experiment, data regarding the listening habits and prior experience of the participants were collected by means of a questionnaire to ensure that the subject were naive[2] listeners.

The absolute category rating method was used [6]. The experimental question for the audiovisual experiment was: "Rate the quality of this loudspeaker" and the anchors were "bad" and "excellent". A discrete 9 point rating scale was used. This experimental question was chosen in order to determine the degree to which visual bias influences audio perception and to allow subjects to evaluate the stimuli as a whole entity without targeting their focus to specific aspects. The rating scale and anchors were inspired by ITU recommendations [4] and [5]. A screenshot of the question and rating scale is shown in figure 2.

The same experimental question was used for the audio-only and visual-only ratings. During the audio-only and the visual-only experiments subjects were allowed to give answers influenced by the audiovisual experiment (since they might have had associated a audio stimuli with a visual stimuli).

To avoid any possible imbalance between modalities caused by the use of degraded audio stimuli, this experiment featured only unprocessed (non-degraded) music excerpts. 5 excerpts were selected to cover a wide range of music genres, having different frequency contents and dynamics. The music excerpts were selected from commercial and reference recordings and transferred (ripped) to a computer (44.1 kHz, 16 bit). The excerpts were carefully selected to include a complete musical phrase, their duration ranging from 9 to 13 sec. The excerpts were equalized with respect to loudness using a loudness model by Moore [8], [9]. Stimuli were presented at comfortable listening levels through circumaural headphones (BeyerDynamic DT 990 PRO).

The 5 music excerpts used in the experiment were:

α) A reference classical recording [10]. Classical symphony orchestra performing at high dynamics.

β) A country music recording [11]. The selection is the song's chorus with male vocals, strumming acoustic guitar, snare drum, bass and handclaps.

γ) A reggae recording with a strong bass line [12] that features drums, bass, guitar, keyboards but no vocals.

δ) An up-tempo hard rock recording [13] with male vocals, guitar with a distortion effect, bass and drums.

ϵ) A pop recording [14]. The recording contains the main theme of the song which is made up of several layers of electronically synthesized instruments that resemble drums, piano, bass and strings.

In a pilot experiment [1], 12 loudspeaker models were evaluated and 5 models were judged to be quite different from one another. These 5 loudspeakers models were presented in this experiment as photographs (see figure 3) displayed on a touch screen (ELO Touchsystems ETL12IC, diagonal size: 12 inches)

[2] In the context of this paper, a naive subject is one that has no prior experience from viewing or listening experiments and has limited knowledge of loudspeakers (technical and commercial).

in front of the subjects. The size of the photographs displayed on the screen was 23 x 16.5 cm. The photographs were taken in a controlled manner and they portray an accurate scale of the original. The selected loudspeakers were:

1. Satellite (of a surround system) 1-way unit in grey plastic cabinet. Dimensions: 12.5 x 9 cm. Diaphragm not visible.
2. Large bookshelf 3-way loudspeaker with 4:3 cabinet proportions. Dimensions: 29 x 41 cm. Diaphragm not visible.
3. Large bookshelf 2-way unit with a rectangular wooden cabinet. Dimensions: 35 x 23 cm. Diaphragm visible.
4. Floor standing 4-way loudspeaker. Dimensions: 184 x 18.5 cm. Diaphragm not visible.
5. Small bookshelf 1-way unit in black plastic cabinet with a tilted upper section. Dimensions: 20.5 x 13 cm. Diaphragm not visible.

In previous studies [1], each subject underwent extensive familiarization prior to the actual experiment. In 2 different familiarization sessions all visual and then all audio stimuli were presented in isolation. A third session featured selected audiovisual combinations presented in the same way as in the actual experiment.

The independent presentation of audio and visual stimuli could have impaired the perception of unity in the audiovisual combinations. Therefore, in this study only a single familiarization featuring audiovisual stimuli was given. This familiarization was identical to the audiovisual experiment and consisted of 2 trials featuring 2 loudspeaker photographs and 2 music excerpts (others than the ones used in the actual experiment).

2.2. Setup

The experiment was conducted in a laboratory room conforming with the ITU-R Rec. BS.1116 recommendation for multichannel listening rooms [15] and controlled from an adjacent control room.

The experiment was fully automated and controlled by a PC running custom-made software (programmed in Labview 6.1). This included the order of presentation of the stimuli (randomization),

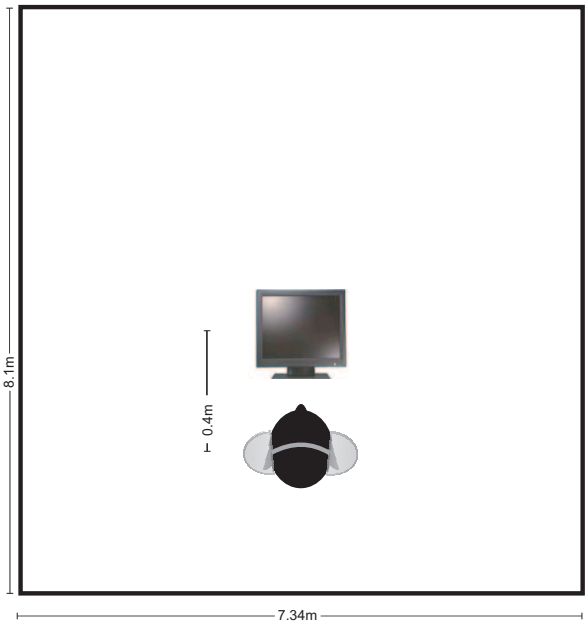


Fig. 4: Diagram of the setup.

Fig. 5: Photograph of the setup.

Fig. 3: The loudspeaker photographs featured in this study. From left to right the visual stimuli are A, B, C, D, E. For all non-floor standing loudspeakers the photograph was manipulated to include the same loudspeaker stand.

the stimuli generation and the data collection. The audio stimuli were generated by an internal sound card (RME Digi 9636, 24-bit, 96 kHz), were converted to analog (Tracer Technologies Big Daddi, 24 bit D/A converter), fed to a headphone amplifier (Behringer Powerplay Pro-XL HA4700) and reproduced by headphones (Beyerdynamic DT 990 PRO circumaural, diffuse field equalized headphones [3]).

The listening room was kept completely dark. The only source of light in the room was the touch screen in front of the subjects.

The light and viewing conditions in the laboratory room are shown in table 1.

The headphones were calibrated to produce a 75dB ± 1.3dB SPL at the listening position when reproducing 1/3 octave band-limited pink noise with center frequency either at 400 or 1000 Hz (-6dBFS at 44100 Hz, 16 bit), measured with a head and torso simulator (B&K HATS 4100, with B&K 4190 microphones and B&K 2669 preamplifiers, with an overall frequency range 6Hz - 20kHz).

3. RESULTS

The evaluations obtained by the subjects for the audiovisual experiment are shown as Latin Squares in figure 1. The 4 factors are *subjects*, *audio*, *visual* and *repetitions*. The levels of factor *subjects* are represented as Latin numerals, the audio stimuli as lowercase Greek letters, the visual stimuli as uppercase Latin letters and the repetitions as #1 and #2.

The averages for each level of each factor are shown in table 2. Audio stimuli 1 and 4 are evaluated to be somewhere in the middle of the rating scale, whereas stimuli 2, 3 and 5 are on the higher end of the scale. Between the 2 groups there is a difference of 1.5 points, and the largest difference between stimuli is 2.6 points. The 2 lowest ranking audio stimuli are a classical and a hard rock recording that do not share common features. For factor *visual*, stimulus 3 is shown to be rated lower than the other stimuli, the minimum and maximum differences between the 2 groups being 1.1 and 1.8 points. The lowest ranking visual stimulus is a typical 2-way loudspeaker that might seem too conventional or old fashioned. The highest ranking visual stimulus is the smallest in size loudspeaker.

The ANOVA table is shown in table 3. The ANOVA analysis suggests that factor *subjects* has negligible influence to the results, while factor *visual* and factor *audio* are influential to the results. The Sum of Squares column shows that factor *audio* is the largest source of variation. The ratio of the Sum of Squares of each factor to the total Sum of Squares can be expressed as a percentage contribution of each factor to the ANOVA model (see [3] page 234). For this experiment, factor *audio* accounts for 45% of the variability of the experiment, while factor *visual* accounts for 16%.

[3] These are according to the manufacturer diffuse field equalized headphones. ITU Rec. BS.1116 [15] recommends diffuse field equalization for headphones for subjective evaluations.

Table 1: Light and viewing conditions in the laboratory room. Luminance measured with Gossen MAVOLUX 5032C, Class C acc. DIN 5032-7, Min. Sensitivity: 0.1 lx.

Background room illumination	0 lx
Listening room dimensions	$8.1 * 7.34 * 2.86$ m
Viewing distance to the touch screen	0.45 m

Table 2: The averages for each level for each factor in the audiovisual experiment. The grand average (g.a.) is 6.32.

Averages			
subjects	audio	visual	repetition
I: 6.8	α: 5.2	A: 7.0	# 1: 6.56
II: 6.4	β: 7.6	B: 6.6	# 2: 6.08
III: 5.9	γ: 7.0	C: 5.2	
IV: 6.3	δ: 5.0	D: 6.5	
V: 6.2	ϵ: 6.8	E: 6.3	
grand average: 6.32			

Table 3: ANOVA table for the audiovisual experiment.

ANOVA table					
Source of Variance	Sum of Squares	Degrees of Freedom	Mean Squares	Ratio of Mean Squares	Significance Probability P
subjects	4.28	4	1.07	$F_{4,36}$=1.01	0.4
audio	53.28	4	13.32	$F_{4,36}$=12.566	<0.0001
visual	18.28	4	4.57	$F_{4,36}$=4.31	0.006
repetitions	2.88	1	2.88	$F_{1,36}$=2.716	0.108
residuals	38.16	36	1.06		
Total (deviations from g.a.)	116.88	49			

Table 4: Average ratings for the audio-only experiment.

audio									
subject		1	2	3	4	5		subject averages	subject deviations
I		7.5	8.5	8.5	8.5	7.5		8.1	1.24
II		5.5	6.5	6.5	4.0	6.0		5.7	-1.16
III		6.0	7.0	8.0	6.0	7.0		6.8	-0.06
IV		5.5	9.0	7.5	5.5	7.5		7.0	0.14
V		5.5	7.0	7.0	6.5	7.5		6.7	-0.16
audio averages		5.9	7.6	7.5	6.1	7.1			
audio deviations		-0.86	0.74	0.64	-0.76	0.24			
grand average: 6.86									

Table 5: ANOVA table of the audio-only experiment.

ANOVA table					
Source of Variance	Sum of Squares	Degrees of Freedom	Mean Squares	Ratio of Mean Squares	Significance Probability P
subjects	29.32	4	7.33	$F_{4,41}$=9.58	0
audio	23.32	4	5.83	$F_{4,41}$=7.62	0.0001
residuals	31.38	41	0.76537		
Total (deviations from g.a.)	84.02	49			

Table 6: Average ratings for the visual-only experiment.

visual									
subject		1	2	3	4	5		subject averages	subject deviations
I		7.0	6.0	5.0	8.0	7.0		6.6	0.82
II		3.5	5.5	7.5	7.0	4.5		5.6	-0.18
III		3.0	7.0	5.0	6.0	6.0		5.4	-0.38
IV		7.0	6.0	3.0	8.0	7.0		6.2	0.42
V		6.0	4.0	3.0	9.0	3.5		5.1	-0.68
visual averages		5.3	5.7	4.7	7.6	5.6			
visual deviations		-0.48	-0.08	-1.08	1.82	-0.18			
grand average: 5.78									

Table 7: ANOVA table of the visual-only experiment.

ANOVA table					
Source of Variance	Sum of Squares	Degrees of Freedom	Mean Squares	Ratio of Mean Squares	Significance Probability P
subjects	14.88	4	3.72	$F_{4,41}$=1.81	0.1452
visual	47.48	4	11.87	$F_{4,41}$=5.78	0.0009
residuals	84.22	41	2.05		
Total (deviations from g.a.)	146.58	49			

Statistical analysis showed that the data follow a normal distribution.

Results for the audio-only experiment are shown in table 4, together with the averages and grand average. Each result represents the average of 2 evaluations for a given stimuli and subject. The ANOVA table is shown in table 5. The ANOVA analysis shows that both factors *audio* and *subjects* are influential to the results. This outcome for factor *audio* is consistent with the results of the audiovisual experiment, again with the same 2 groups being different (audio stimuli 1 and 4 rated lower than the rest). However the differences between levels are smaller than in the audiovisual experiment and not greater than 1.7 rating scale points. Subjects is a nuisance factor that should not be influential but in this case there is 1 subject giving answers in the low end of the rating scale and another in the high end of the rating scale. The largest difference is 2.4 points. Differences like that are to be expected since no reference is used in the experiment. Additionally there is only a small number of subjects in this study and this makes such differences more prominent.

Results for the visual-only experiment are shown in table 6, together with the averages and grand average. Each result represents the average of 2 evaluations for a given stimuli and subject. The ANOVA table is shown in table 7. The ANOVA analysis shows that factor *visual* is influential to the results. The average ratings show 3 stimuli with similar ratings, one stimulus with a lower rating and one stimulus that is rated much higher than the rest. The largest difference is 2.9 rating scale points, which in the context of this experiment is very important.

3.1. Data across experiments

The ranking of audio and visual stimuli for the audiovisual as well as for the audio-only and visual-only experiments is shown in table 8. The audio stimuli are almost identically ranked in the audiovisual and audio-only experiments (there is an inversion between stimuli 1 and 4), while the visual stimuli ranking between experiments is different. Figures 6 and 7 show the means ± standard deviations of the data in the 3 experiments. Close resemblance is seen for the audio ratings in the audiovisual and audio-only experiments. The visual ratings between the audiovisual and visual-only experiments

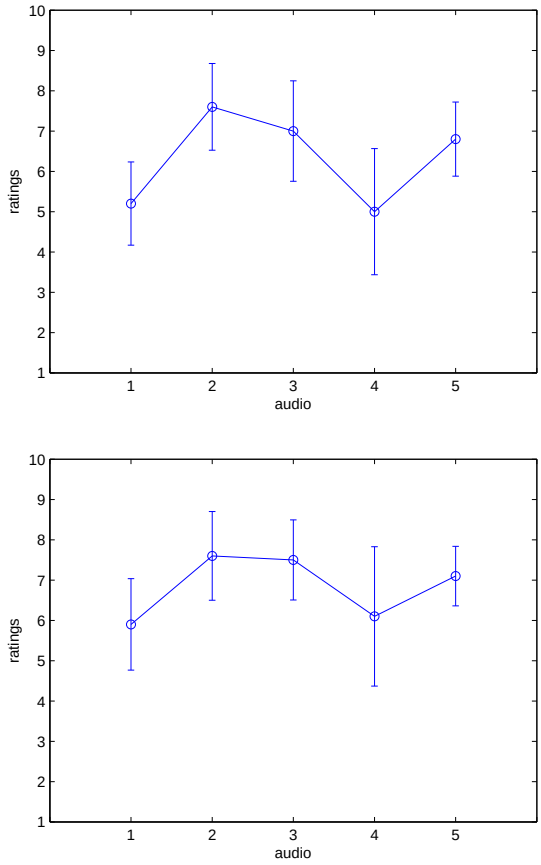


Fig. 6: Plots of the mean and ± 1 standard deviation for the audio stimuli of the audiovisual (top plot) and audio-only (bottom plot) experiments.

exhibit differences. The highest ranking visual stimulus in the audiovisual experiment (the smallest in size loudspeaker) is rated second lowest in the visual-only experiment. The highest ranking visual stimuli in the visual-only experiment is the largest in size loudspeaker and was ranked 3^{rd} in the audiovisual experiment. However, stimulus 3 (a medium sized 2-way loudspeaker) is rated lowest in both cases.

Interestingly the maximum difference between audio stimuli in the A(AV) results is 2.6 points and 1.7 points in the audio-only results. Furthermore, the audio-only ratings are overall higher than the A(AV) ratings and the visual-only ratings are lower than the V(AV) ratings (with the exception of stimuli 4 that in the visual-only case was rated highest).

Table 8: Data across experiments, averaged across subjects. The ranking order is shown from lowest to highest. A(AV) and V(AV) are the averaged results for the audiovisual experiment with respect to the audio and visual stimuli respectively.

experiment	A(AV)	V(AV)	A-only	V-only
ranking	4,1,5,3,2	3,5,4,2,1	1,4,5,3,2	3,1,5,2,4

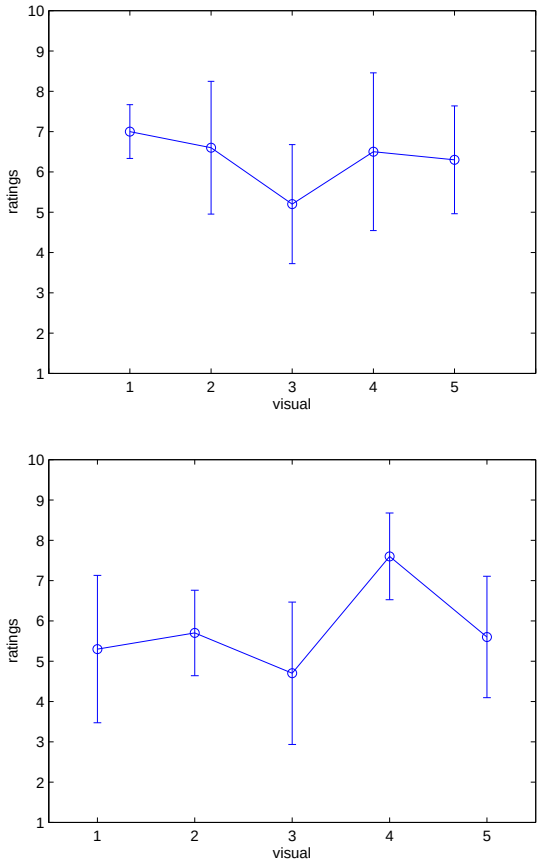


Fig. 7: Plots of the mean and ± 1 standard deviation for the visual stimuli of the audiovisual (top plot) and visual-only (bottom plot) experiments.

4. CONCLUSIONS

The Latin Square design was efficient and although a small number of subjects participated in the study, the ANOVA was powerful enough to detect the differences between the main factors. These differences are shown to be not only statistically significant but also significant in terms of the context of the experiment.

A disadvantage of the LS design is that it rules out the analysis of interactions between audio stimuli, visual stimuli and the subjects. A previous study [1] with similar stimuli has shown that interactions between these factors have a small but statistically significant influence to the results.

The overall conclusions of this and the previous studies share one common important point: the auditory modality is the dominant source of influence and the overall audiovisual evaluation produces results different than the linear combination of the unimodal experiments. However, in contrast to the previous studies the results presented here show a statistically significant and important influence of the visual modality to the overall evaluation.

5. ACKNOWLEDGEMENTS

The author would like to thank Kasper K. Berthelsen of the Department of Mathematical Sciences, Aalborg University, for fruitful discussions on statistical issues.

6. REFERENCES

[1] Karandreas, A. and Christensen, F., "Influence of visual appearance on loudspeaker sound quality evaluation", 124^{th} Audio Eng. Soc. Conv., Amsterdam, 2008.

[2] Hollier, M.P, and Voelcker, R.M., "Towards a multi-modal perceptual model", BT Technol. J. Vol. 14 No. 4 October 1997.

[3] Montgomery D., "Design and Analysis of Experiments", 5^{th}ed., Wiley, 2001.

[4] ITU-T Rec. P.910, "Subjective Video Quality Assessment Methods for Multimedia Applications". International Telecommunications Union, Geneva, Switzerland, 2008.

[5] ITU-R Rec. BT.500-12, "Methodology for the Subjective Assessment of the Quality of Television Pictures", International Telecommunications Union, Geneva, Switzerland, 2009.

[6] ITU-T Rec. P.911, "Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems", International Telecommunications Union, Geneva, Switzerland, 1998.

[7] ITU-T Rec. J.140, "Subjective picture quality assessment for digital cable television systems", International Telecommunications Union, Geneva, Switzerland, 1998.

[8] Glasberg, B.R. and Moore, B.C.J., "A Model of Loudness Applicable to Time-Varying Sounds", J.Audio.Eng.Soc., Vol.50, No. 5, pp.331-342, 2002.

[9] http://hearing.psychol.cam.ac.uk/Demos/demos.html

[10] EBU Parameters for the Subjective Evaluation of Sound Programme material (PEQS) CD, European Broadcasting Union, Geneva, Switzerland, 2008, Track 14: Bruckner - "Symphony No. 3", Slovenia Philharmonic Orchestra / Gyorgy Gyorivanyi, timing [min.]: 0:10 - 0:20

[11] EBU Sound Quality Assessment Material CD, European Broadcasting Union, Geneva, Switzerland, 2008, Track 70: Eddie Rabbitt - "Early in the morning", timing [min.]: 0:00 - 0:09.

[12] Bob Marley and the Wailers, "Uprising - Coming In From The Cold", Island, 2001, ASIN: B00005A7X0, 44100 Hz, 16 bit, timing [min.]: 0:15 - 0:26.

[13] Queens Of The Stone Age, "Lullabies to Paralyze - Little Sister", Interscope Records, 2005, ASIN: B0007QJ1MK, 44100 Hz, 16bit, timing [min.]: 1:39 - 1:51.

[14] Depeche Mode, "Violator - Enjoy the silence", Reprise / Wea, 1990, ASIN: B000002LK1, 44100, 16bit timing [min.]: 2:00 - 2:13.

[15] ITU-R Rec. BS.1116, "Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems", International Telecommunications Union, Geneva, Switzerland, 1997.

3

Data across experiments

This chapter compares the data across experiments in order to give an overview of the effect of the stimuli presentation and the experimental question. Also, a subset of the data is further examined in order to reveal possible effects that the choice of degraded stimuli might have. The following abbreviations are used for the experiments:

- the experiment in Manuscript A : Exp1
- the experiment in Manuscript B : Exp2
- the experiment in Manuscript C : Exp3
- the 1^{st} experiment in Manuscript D : Exp4a
- the 2^{nd} experiment in Manuscript D : Exp4b
- the experiment in Manuscript E : Exp5

3.1 Ranks, means and standard deviations across experiments

A main goal of this research was to establish a methodology for the presentation of audio and visual stimuli in subjective evaluations. Different experiments presented in this thesis investigate various presentation techniques for the audio and visual stimuli. One of the most important issues is whether it is necessary to use the actual product in the experiments. Results presented in this section show that it is not absolutely necessary to have the actual product available during the test. One way to support this is to compare the audio-only (A-only), visual-only (V-only) data as well as the audio part of the AV data[1] and the visual data of the AV data (designated as A(AV) and V(AV) respectively) for the experiments where actual loudspeakers, large-scale and small-scale photographs of the loudspeakers are used and audio reproduction is via loudspeakers and headphones.

The following figures show means and standard deviations of the A-only, V-only, A(AV) and V(AV) data. The visual stimuli are the same across all experiments (so any comparison between the visual stimuli across experiments is valid). The selected loudspeakers and their respective abbreviations are:

[1]That means excluding the visual factor from the analysis.

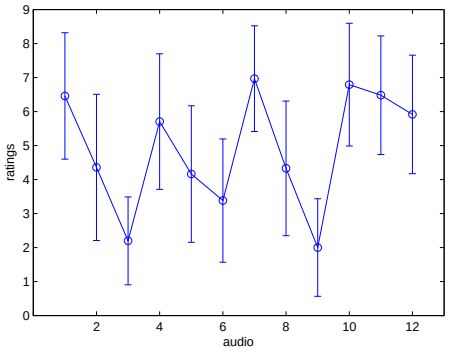
- Satellite (of a surround system) 1-way unit in grey plastic cabinet. Dimensions: 12.5 x 9 cm. Diaphragm not visible.
- Large bookshelf 3-way loudspeaker with 4:3 cabinet proportions. Dimensions: 29 x 41 cm. Diaphragm not visible.
- Large bookshelf 2-way unit with a rectangular wooden cabinet. Dimensions: 35 x 23 cm. Diaphragm visible.
- Floor standing 4-way loudspeaker. Dimensions: 184 x 18.5 cm. Diaphragm not visible.
- Small bookshelf 1-way unit in black plastic cabinet with a tilted upper section. Dimensions: 20.5 x 13 cm. Diaphragm not visible.

Experiments Exp2, Exp3, Exp4a, Exp4b featured 6 degraded versions of a single music excerpt. The excerpt features the chorus of a rock/country recording with male vocals, strumming acoustic guitar, snare drum, bass and handclaps. The excerpt was carefully selected to include a complete musical phrase lasting 9 sec. There were 3 high-pass filtered versions and 3 with added harmonic distortion. The high-pass filtered versions were filtered at 110, 220 and 440 Hz while the harmonically distorted versions were all high-pass filtered at 110 Hz and had added harmonic distortion at 3 distinct levels. The pattern of harmonic distortion was constant and the only difference was the relative level of the harmonic distortion to the 110 Hz high-pass filtered excerpt. Exp5 featured 5 different audio stimuli and should not be compared with the rest of the experiments. Exp1 featured the same audio stimuli (audio stimuli a1 to a6) and 6 degraded versions of another music excerpt (audio stimuli a7 to a12).

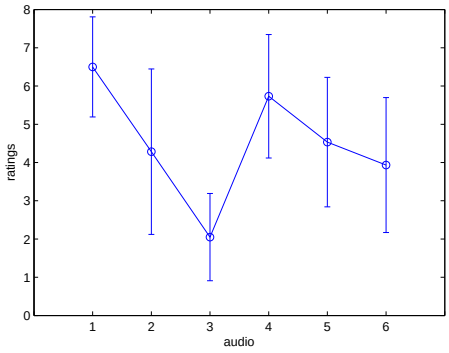
Across all experiments (except Exp5) the A-only and A(AV) data (figures 3.3 and 3.1) exhibit very similar means and standard deviations. For all experiments the V(AV) ratings across the visual levels are very similar (figure 3.2, with the standard deviation across experiments being the only difference (the standard deviation in Exp4b being smaller than for the other experiments). The V-only ratings across experiments show some differences. For Exp1 and Exp3 the V-only ratings are similar. The ratings for Exp2 are similar to Exp1 and Exp3 with the exception that visual stimuli 1 (*v1*) is rated higher (rated 2^{nd} highest in Exp2). For Exp4a the data resembles Exp1, although *v4* is rated lower than Exp1 (in all experiments except Exp4a *v4* is rated highest). In Exp4b *v2* is rated lowest. Furthermore, in Exp1, Exp3 and Exp4a the two smallest in size loudspeakers, *v1* and *v5*, are rated lowest. For Exp5 the results are similar to Exp1 and Exp3, with *v1* and *v4* more elevated.

All in all, these results show that the data collected from these experiments is comparable, and it is thus reasonable to claim that these results support the hypothesis that different stimuli presentations produce equally valid results and that substitutes of an actual product can be be used in subjective evaluations.

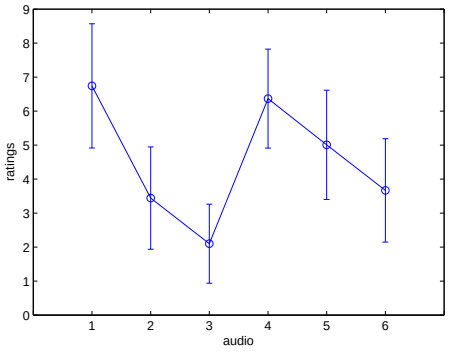
The ranking of the stimuli for all experiments is shown in table 3.1. Across all experiments (except Exp5) the A-only and A(AV) data are very similar, following the same ranking order. The V(AV) ranking across experiments is different. The V-only rankings across experiments are also different, there are however some common points: a) *v4* is ranked highest in all but one experiments. b) Rankings for Exp1 and Exp3 are similar, with a single inversion between *v1* and *v5*. Furthermore, c) *v1* is rated lowest or second-lowest in all but one experiments.



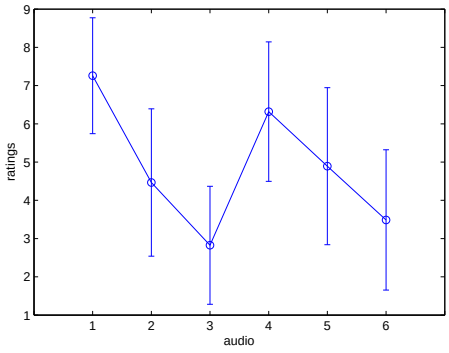
(a) Exp1



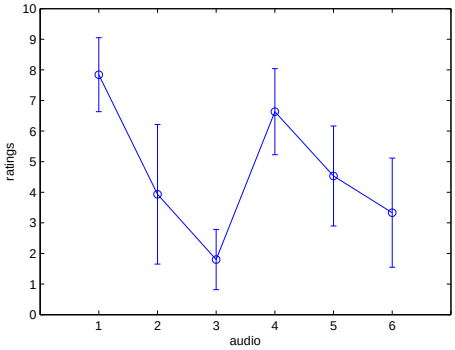
(b) Exp2



(c) Exp3



(d) Exp4a

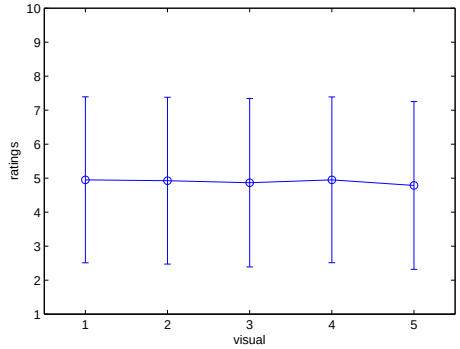


(e) Exp4b

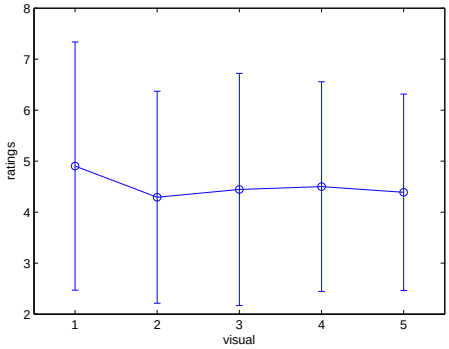
Figure 3.1: A(AV) data from all experiments.

Table 3.1: Ranks across all experiments. The ranking order is shown from lowest to highest. Exp1, only the first 6 audio stimuli are shown because they are the same as the audio in the other experiments. Exp5 features a different set of audio stimuli.

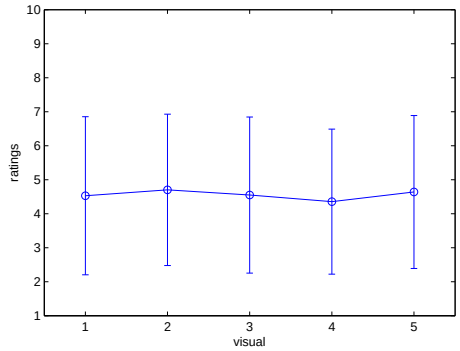
Exp.			test	
	A(AV)	V(AV)	Aonly	Vonly
Exp1	3,6,5,2,4,1	5,3,2,4,1	3,6,2,5,4,1	1,5,3,2,4
Exp2	3,6,2,5,4,1	2,5,3,4,1	3,6,2,5,4,1	5,2,3,1,4
Exp3	3,2,6,5,4,1	4,1,3,5,2	3,2,6,5,1,4	5,1,3,2,4
Exp4a	3,6,2,5,4,1	1,4,5,3,2	3,6,2,5,4,1	1,5,2,4,3
Exp4b	3,6,2,5,4,1	2,3,1,4,5	3,6,2,5,4,1	2,1,5,3,4
Exp5	4,1,5,3,2	3,5,4,2,1	1,4,5,3,2	3,1,5,2,4



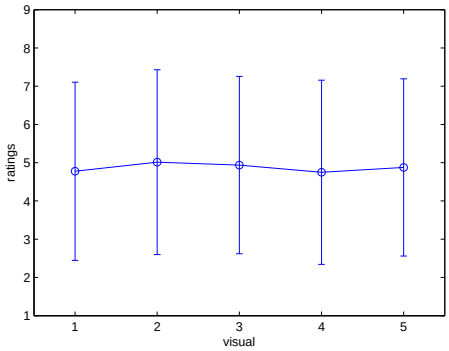
(a) Exp1



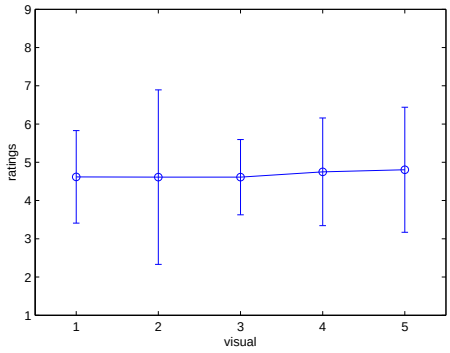
(b) Exp2



(c) Exp3

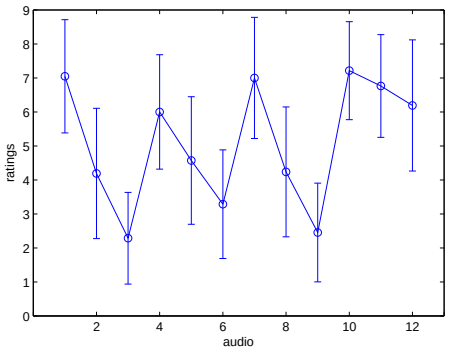


(d) Exp4a

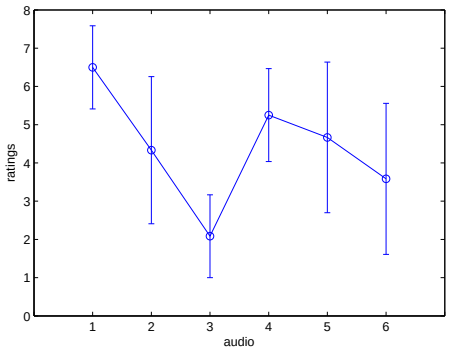


(e) Exp4b

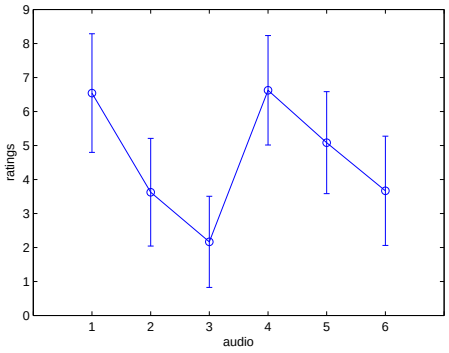
Figure 3.2: V(AV) data from all experiments.



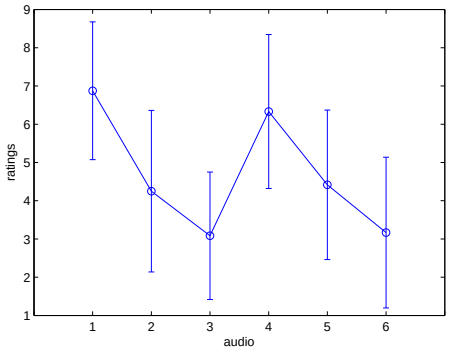
(a) Exp1



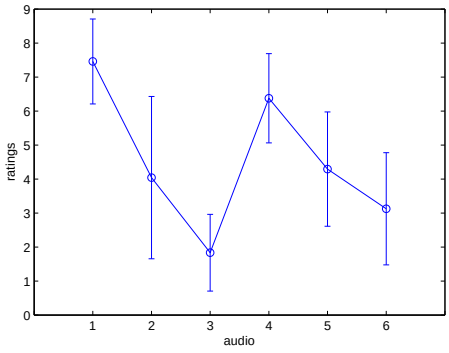
(b) Exp2



(c) Exp3

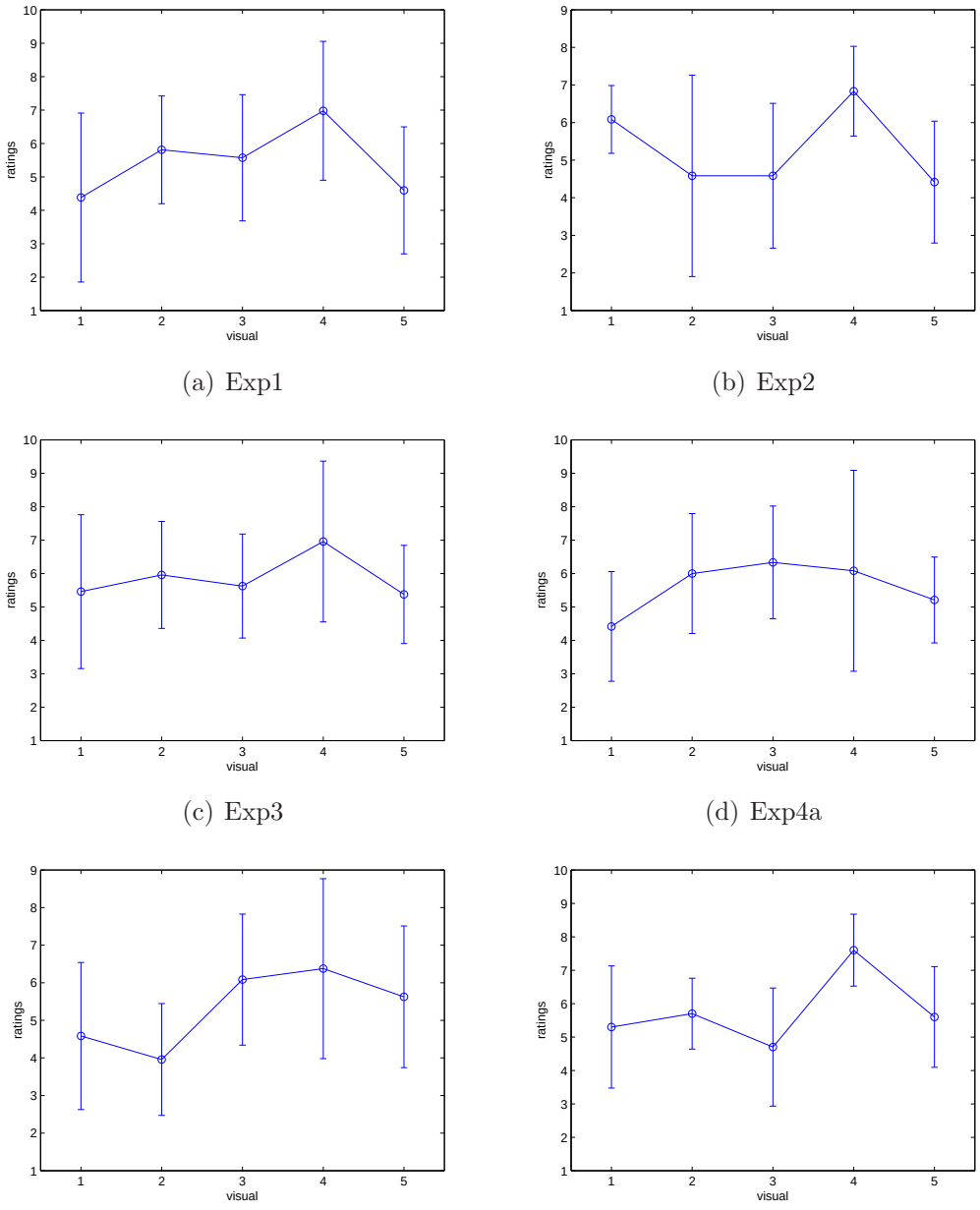


(d) Exp4a



(e) Exp4b

Figure 3.3: A-only data from all experiments.



(a) Exp1 (b) Exp2

(c) Exp3 (d) Exp4a

(e) Exp4b (f) Exp5

Figure 3.4: V-only data from all experiments.

3.2 ANOVA across experiments

This section presents ANOVA across the data of experiments Exp1, Exp2, Exp3, Exp4a and Exp4b. In each comparison the data is pooled and a new factor experiment (*Exp*) is introduced. Each level of factor *Exp* signifies the corresponding experiment. The levels of factor audio (A) and visual (V) represent the audio and visual stimuli irrespective of the presentation technique. The ANOVA tables give information on whether the different stimuli presentation and experimental question influenced the results. Thus, the comparisons presented here are such that only 1 parameter is changed at a time. Presented first are cases where only the visual presentation differed, followed by a case where only the audio presentation was different and finally a case where the difference lies in the experimental question (the last case is presented in more detail in Manuscript D).

For the analysis in this section, only half of the data from Exp1 is used (Exp1 features degradations of two music excerpts while Exp2, Exp3, Exp4a and Exp4b feature degradations of one of the two music excerpts), and Exp5 is excluded. Figures 3.5 and 3.6, show that the selected data set from the Exp1 is representative of the whole data set. The P-values and Sum of Squares values are comparable for all terms.

The ANOVA between Exp1 and Exp2 (figure 3.7) shows the analysis for 2 experiments whose only difference is the visual presentation. The analysis shows that factor *audio* is the only statistically significant term. Factor *Exp* is a 2 level factor where each level stands for either Exp1 or Exp2. Factor *Exp* is not statistically significant showing that the stimuli presentation did not influence the results.

The ANOVA between Exp1 and Exp3 (the only difference among the 2 experiments is the presentation of visual stimuli) is shown in figure 3.8. The analysis shows that factor *audio*, factor *Exp* and the *A*Exp* interaction are statistically significant, showing that in this case the stimuli presentation did influence results. However, neither factor *visual* or the *V*Exp* interaction are statistically significant, showing that the influence of the visual stimuli on the overall audiovisual perception is unaltered.

Analysis of Variance

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
S	1255.4	20	62.772	36.59	0
A	7065.3	11	642.303	374.42	0
V	10.2	4	2.557	1.49	0.2028
S*A	2714.9	220	12.34	7.19	0
S*V	185.9	80	2.324	1.35	0.0231
A*V	138.2	44	3.14	1.83	0.0009
S*A*V	1637.7	880	1.861	1.08	0.0939
Error	2161.5	1260	1.715		
Total	15169.1	2519			

Constrained (Type III) sums of squares.

Figure 3.5: Analysis for Exp1. Both music excerpts are included in this analysis. S stands for factor *subjects*, A for *audio* and V for *visual*.

Analysis of Variance

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
S	1321.12	20	66.056	39.19	0
A	2496.65	5	499.33	296.21	0
V	9.14	4	2.285	1.36	0.2481
S*A	956.52	100	9.565	5.67	0
S*V	190.26	80	2.378	1.41	0.0145
A*V	59.71	20	2.985	1.77	0.0204
S*A*V	800.29	400	2.001	1.19	0.028
Error	1062	630	1.686		
Total	6895.69	1259			

Constrained (Type III) sums of squares.

Figure 3.6: Analysis for Exp1. Only music excerpt 1 is included in this analysis.

Analysis of Variance

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
A	2202.24	5	440.449	130.91	0
V	17.56	4	4.39	1.3	0.2661
Exp	4.69	1	4.686	1.39	0.2381
A*V	37.02	20	1.851	0.55	0.9455
A*Exp	17.35	5	3.471	1.03	0.3973
V*Exp	11.24	4	2.811	0.84	0.5025
A*V*Exp	30.27	20	1.513	0.45	0.9827
Error	5248.52	1560	3.364		
Total	8574.36	1619			

Constrained (Type III) sums of squares.

Figure 3.7: Analysis for Exp1 (excerpt 1 data) and Exp2.

Analysis of Variance

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
A	4158.2	5	831.646	267.59	0
V	0.8	4	0.196	0.06	0.9927
Exp	14.5	1	14.513	4.67	0.0308
A*V	29.2	20	1.462	0.47	0.9774
A*Exp	150.8	5	30.152	9.7	0
V*Exp	18.6	4	4.654	1.5	0.2004
A*V*Exp	31.8	20	1.59	0.51	0.9633
Error	5967.2	1920	3.108		
Total	10530.1	1979			

Constrained (Type III) sums of squares.

Figure 3.8: Analysis for Exp1 (excerpt 1 data) and Exp3.

Analysis of Variance

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
A	2199.27	5	439.855	175.57	0
V	9.1	4	2.276	0.91	0.4582
Exp	0.57	1	0.567	0.23	0.6343
A*V	18.2	20	0.91	0.36	0.9956
A*Exp	58.12	5	11.624	4.64	0.0003
V*Exp	18.78	4	4.696	1.87	0.1127
A*V*Exp	25.96	20	1.298	0.52	0.9603
Error	2555.37	1020	2.505		
Total	5294.44	1079			

Constrained (Type III) sums of squares.

Figure 3.9: Analysis for Exp2 and Exp3.

For Exp2 and Exp3 (figure 3.9) only factor *audio* and interaction *A*Exp* are statistically significant, with the visual stimuli presentation having a small and not statistically significant effect.

An overall analysis incorporating data from the 3 aforementioned experiments, where the visual stimuli presentation differed but the audio presentation was unaltered, is shown in figure 3.10. Factor *audio* and interaction *A*Exp* are statistically significant, and factor *Exp* is nearly statistically significant. Factor visual has a small, not statistically significant effect.

The audio presentation was the only difference between Exp3 and Exp4a. The ANOVA is shown in figure 3.11. Factor *Exp* is statistically significant. The interaction *A*Exp* shows that there were differences in the influence of the audio stimuli in each of the 2 experiments, that could be attributed to the stimuli presentation.

In Manuscript D a comparison between Exp4a and Exp4b is discussed, where the difference lies in the experimental question. The ANOVA comparing the 2 experiments is shown in figure 3.12. Factor *Exp* is statistically significant. The interaction *A*Exp* shows that there were differences in the audio evaluations among

Analysis of Variance

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
A	3915.1	5	783.011	255.87	0
V	7.5	4	1.865	0.61	0.6559
Exp	15.8	2	7.893	2.58	0.0761
A*V	26.7	20	1.335	0.44	0.9857
A*Exp	166.4	10	16.641	5.44	0
V*Exp	32.2	8	4.026	1.32	0.2309
A*V*Exp	59.8	40	1.496	0.49	0.9972
Error	6885.6	2250	3.06		
Total	12205.3	2339			

Constrained (Type III) sums of squares.

Figure 3.10: Analysis for Exp1 (excerpt 1 data), Exp2 and Exp3.

Analysis of Variance

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
A	3563.41	5	712.682	253.07	0
V	13.99	4	3.498	1.24	0.291
Exp	36.52	1	36.521	12.97	0.0003
A*V	23.99	20	1.2	0.43	0.9876
A*Exp	77.04	5	15.409	5.47	0.0001
V*Exp	1.78	4	0.444	0.16	0.9596
A*V*Exp	30.83	20	1.542	0.55	0.9468
Error	3886.26	1380	2.816		
Total	7638.98	1439			

Constrained (Type III) sums of squares.

Figure 3.11: Analysis for Exp3 and Exp4a.

Analysis of Variance

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
A	4519.03	5	903.805	307.8	0
V	3.15	4	0.788	0.27	0.8983
Exp	13.48	1	13.481	4.59	0.0323
A*V	32.91	20	1.646	0.56	0.9398
A*Exp	102.47	5	20.495	6.98	0
V*Exp	7.48	4	1.871	0.64	0.6361
A*V*Exp	41.3	20	2.065	0.7	0.8261
Error	4052.09	1380	2.936		
Total	8777.1	1439			

Constrained (Type III) sums of squares.

Figure 3.12: Analysis for Exp4a and Exp4b.

the 2 experiments that could be attributed to the experimental question.

3.3 The effect of audio degradation on the AV evaluation

There are 3 main differences between Exp5 and the other experiments: 1) the audio stimuli are not degraded 2) the experimental question is neutral and 3) the usage of a Latin Square design which among other things means that subjects are presented with unique AV combinations. These differences could result in minimizing possible biases and allow subjects to focus equally to the audio and visual stimuli. In the other experiments it is possible that the subjects focus their attention towards audio because they have to assess the additional effect of degradation of the audio stimuli when at the same time there is no degradation in the visual stimuli. It is interesting for those experiments to isolate the AV data that feature the least degraded audio excerpt and repeat the ANOVA in order to establish whether the influence of the visual modality is stronger for these AV evaluations [2].

In the following analysis the least degraded audio data is isolated so that any differences between the AV ratings should be caused by the influence of factors *visual* and *subjects*. The analysis for each experiment is shown in ANOVA tables, plots of means and standard deviations and plots showing lower quartile, median, and upper quartile values. The ANOVA analysis shows that there is a significant effect for factor *visual* in Exp4b. For all other experiments the ANOVA results show that there are no significant differences between the levels of factor *visual* for the AV data that features the least degraded audio excerpt.

[2](Zielinski, Rumsey and Bech, 2003) investigated this topic: *"Another interesting issue related to the interaction between visual and audio modalities is the hypothesis that video presence may affect the evaluation of audio quality for slightly impaired items only. In other words, it was hypothesized that the video presence may fix some quality imperfections for the least impaired items whereas severely impaired items might be too bad to be fixed. To check this hypothesis, the ANOVA test was repeated for selected items having the least degraded quality. Results of this analysis did not reveal any significant difference from the results obtained previously for all items, and therefore this hypothesis was rejected".*

Analysis of Variance

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
V	11.95	4	2.9875	2.79	0.0308
S	49.142	5	9.82833	9.19	0
V*S	16.65	20	0.8325	0.78	0.732
Error	96.25	90	1.06944		
Total	173.992	119			

Constrained (Type III) sums of squares.

Figure 3.13: Analysis for the least degraded audio for Exp4b.

Analysis of Variance

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
V	2.367	4	0.5917	0.37	0.8318
S	100.742	5	20.1483	12.48	0
V*S	24.633	20	1.2317	0.76	0.7493
Error	145.25	90	1.6139		
Total	272.992	119			

Constrained (Type III) sums of squares.

Figure 3.14: Analysis for the least degraded audio for Exp4a.

Analysis of Variance

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
V	4.117	4	1.0292	0.88	0.4792
S	270.442	5	54.0883	46.25	0
V*S	17.183	20	0.8592	0.73	0.7804
Error	105.25	90	1.1694		
Total	396.992	119			

Constrained (Type III) sums of squares.

Figure 3.15: Analysis for the least degraded audio for Exp3.

Analysis of Variance

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
V	11.3333	4	2.83333	1.77	0.1514
S	6.1	2	3.05	1.91	0.1604
V*S	11.5667	8	1.44583	0.9	0.5219
Error	72	45	1.6		
Total	101	59			

Constrained (Type III) sums of squares.

Figure 3.16: Analysis for the least degraded audio for Exp2.

Analysis of Variance

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
V	11.067	4	2.7667	1.27	0.2849
S	318.514	20	15.9257	7.33	0
V*S	164.533	80	2.0567	0.95	0.598
Error	228	105	2.1714		
Total	722.114	209			

Constrained (Type III) sums of squares.

Figure 3.17: Analysis for the least degraded audio for Exp1.

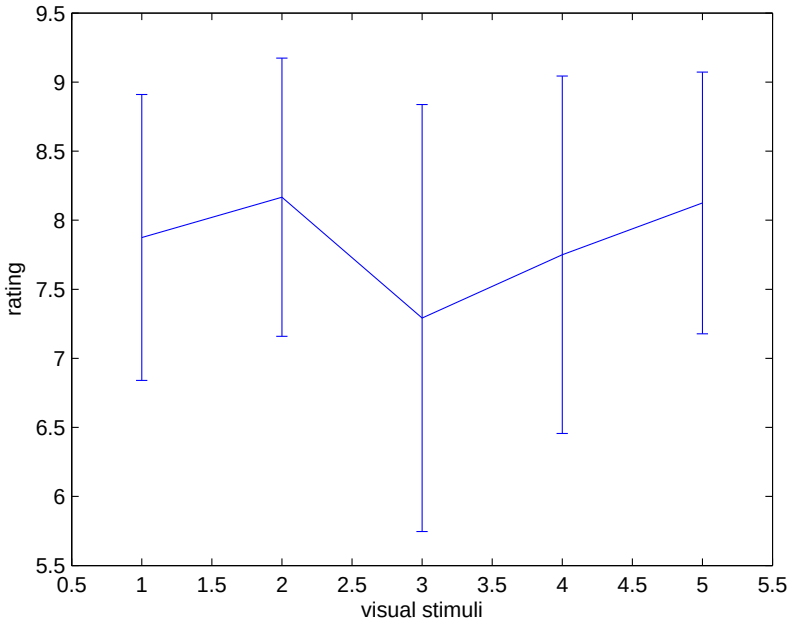


Figure 3.18: Analysis for the least degraded audio for Exp4b.

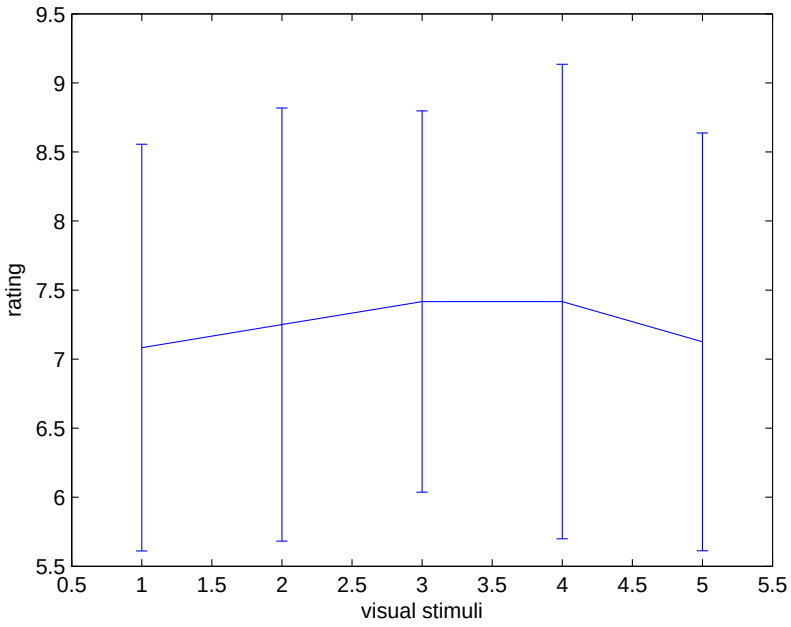


Figure 3.19: Analysis for the least degraded audio for Exp4a.

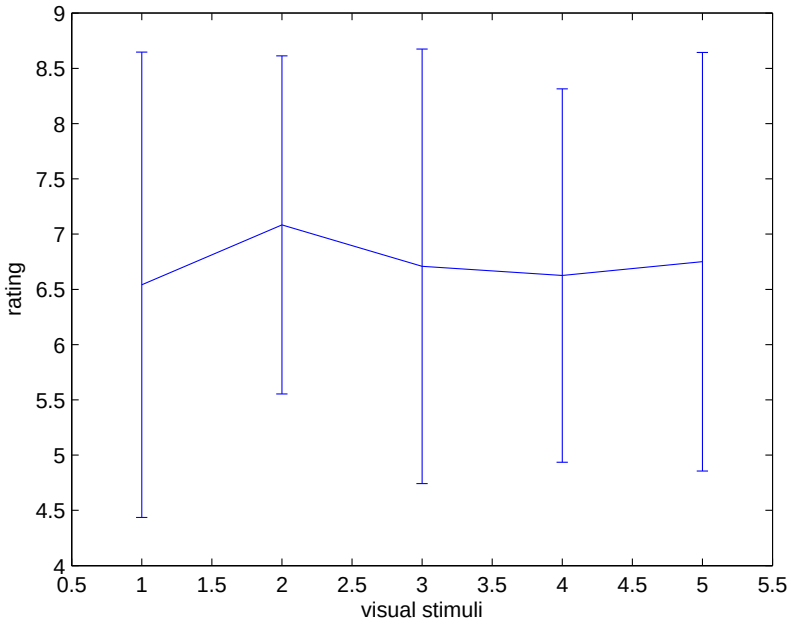


Figure 3.20: Analysis for the least degraded audio for Exp3.

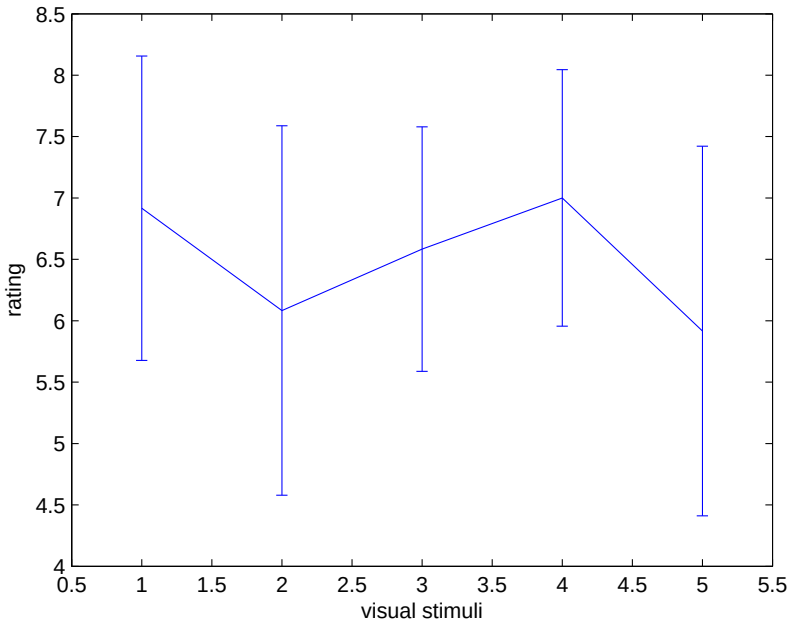


Figure 3.21: Analysis for the least degraded audio for Exp2.

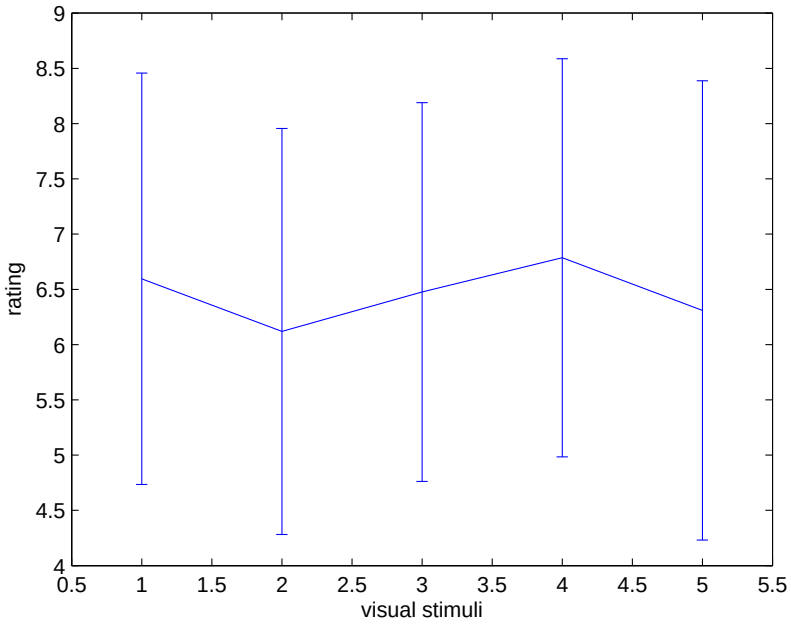


Figure 3.22: Analysis for the least degraded audio for Exp1.

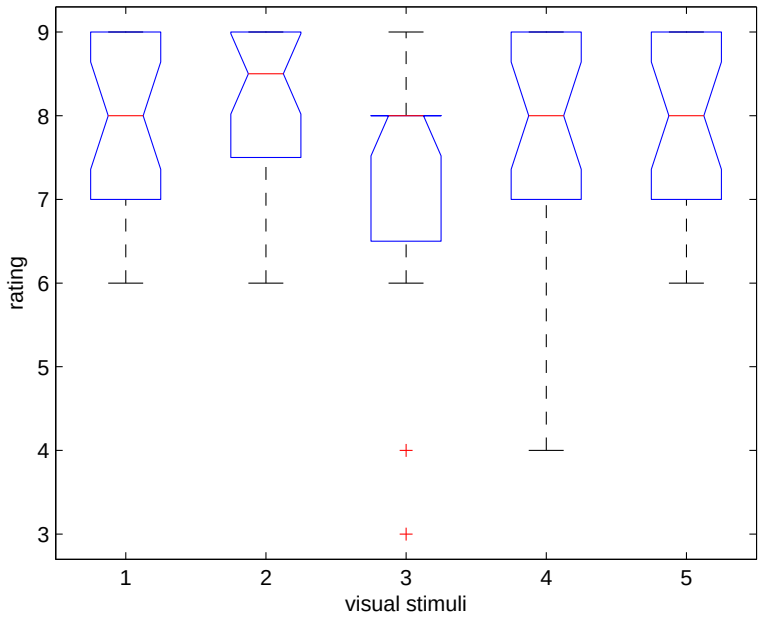


Figure 3.23: Analysis for the least degraded audio for Exp4b.

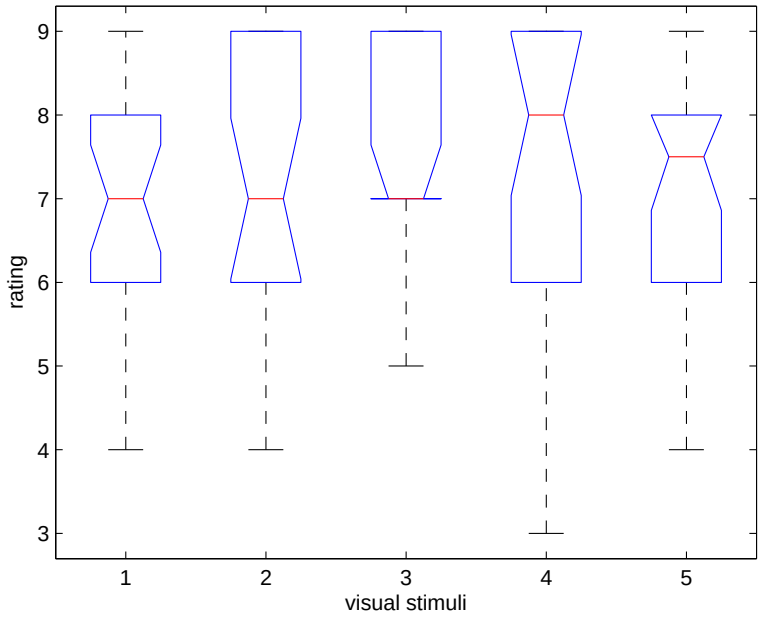


Figure 3.24: Analysis for the least degraded audio for Exp4a.

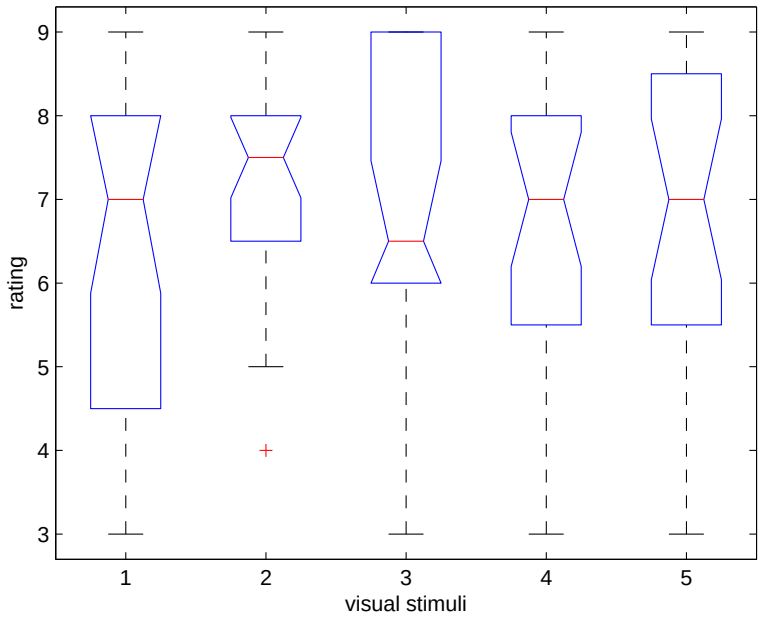


Figure 3.25: Analysis for the least degraded audio for Exp3.

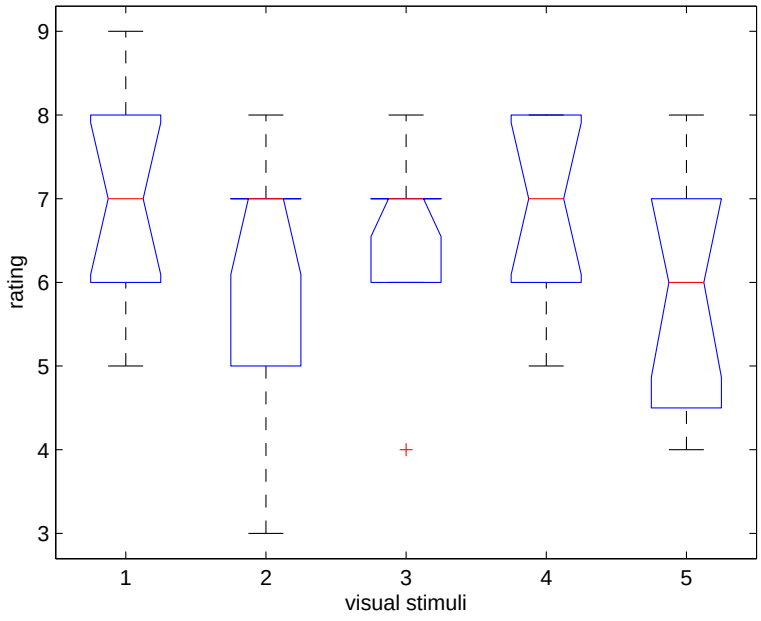


Figure 3.26: Analysis for the least degraded audio for Exp2.

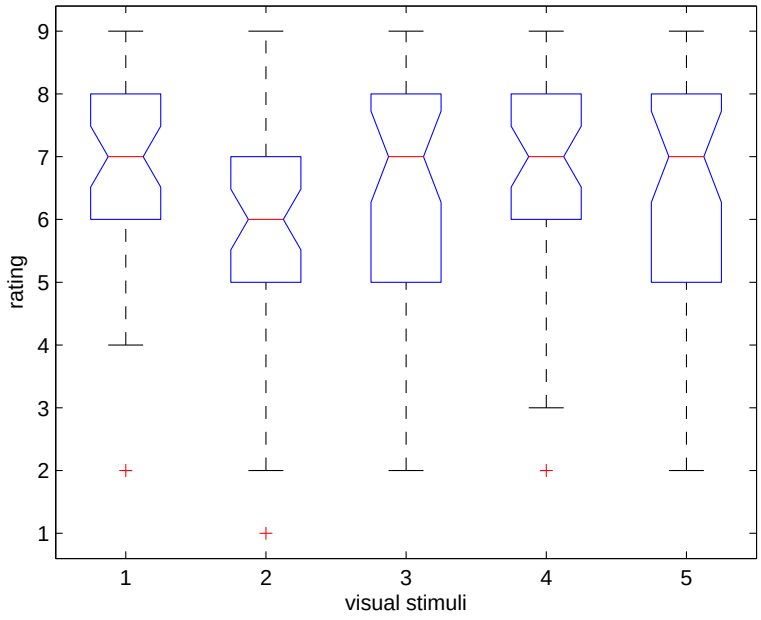


Figure 3.27: Analysis for the least degraded audio for Exp1.

4

General conclusions

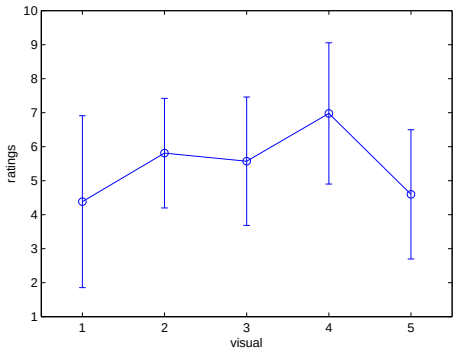
Manuscript A described the selection of 5 out of 12 visual stimuli. The reasoning for the selection process was to include visual stimuli that would cover the whole range of the rating scale. An important reason for limiting the number of loudspeakers was that it was impractical to have 12 loudspeaker pairs as part of the setup due to their size, and made their presentation very cumbersome (since each visual object should be presented independently). It was thus important to compare the usage of an actual product and a substitute. The results of the experiments described across manuscripts A to D show that substitutes can be effectively used in the subjective evaluation of audiovisual products. These substitutes are of course easier to handle and thus make the setup much more practical and easy to implement.

According to a popular notion, loudspeaker size and loudspeaker appearance affect a buyer's choice. The V-only plots across experiments (figure 4.1) show a pattern that suggests that the smallest loudspeakers are rated lowest while the largest are rated highest. This was also shown in the initial pilot test, used to select the visual stimuli where 12 loudspeakers were evaluated in manuscript A (presented in this section in figures 4.2 and 4.3). The order of size is similar to the quality rank reported by the subjects (see table 4). The fact that size is not shown to be highly correlated to the quality rankings in the AV test might be because *audio* dominates over *visual* for the specific context.

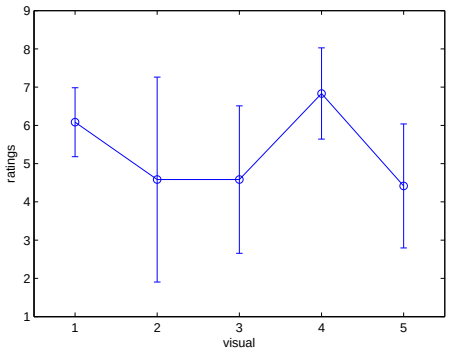
In this work subjects were not asked to rate stimuli on the grounds of which loudspeaker they would choose for their homes or which they would buy. It is possible that for practical or psychological reasons, the same subjects might have shown different preference for the same set of loudspeakers if they considered which purchase to make, and this decision might not necessarily be due to price but considerations like the appropriateness of the loudspeaker for their own needs (style/design,

Table 4.1: Comparison of the size and ranking of the loudspeakers used in the pilot test in manuscript A (selection of visual stimuli).

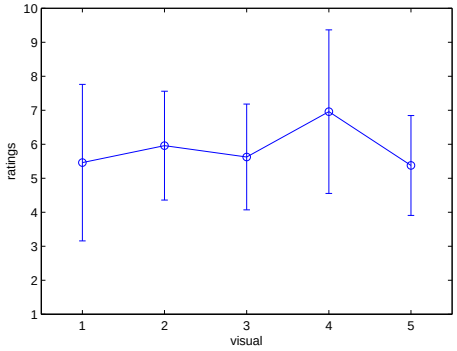
visual stimuli according to Manuscript A	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12
size (smallest to largest)	2, 7, 3, 12, 5, 9, 8, 10, 4, 6, 1, 11
ranking (lowest to highest)	2, 3, 5, 7, 12, 10, 8, 6, 9, 4, 1, 11
stimuli v1-v5 correspond to:	2, 4, 10, 11, 7



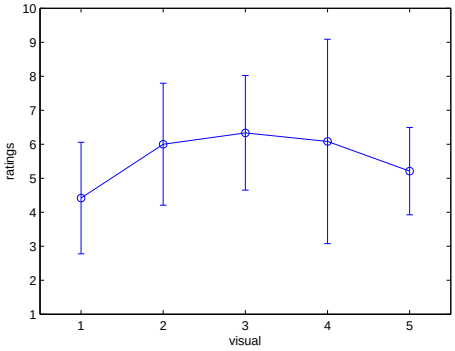
(a) Exp1



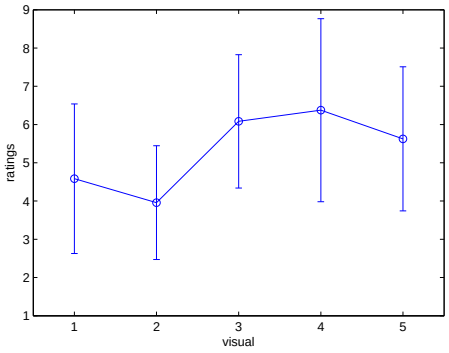
(b) Exp2



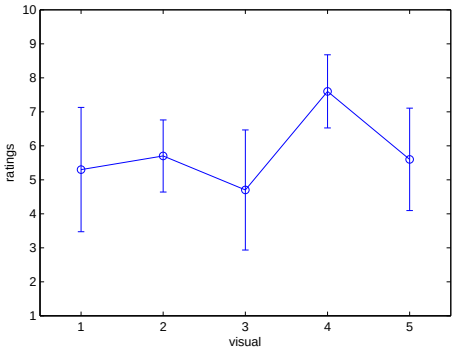
(c) Exp3



(d) Exp4a



(e) Exp4b



(f) Exp5

Figure 4.1: V-only data from all experiments. Stimuli *v1* and *v5* are the smallest in size, *v2 v3* are intermediate and *v4* is the largest.

Figure 4.2: An ensemble of all the loudspeakers used in the pilot test, showing the range of loudspeakers used. Aspect ratio is not maintained in this figure. The range of height covered is from 12.5cm to 184cm.

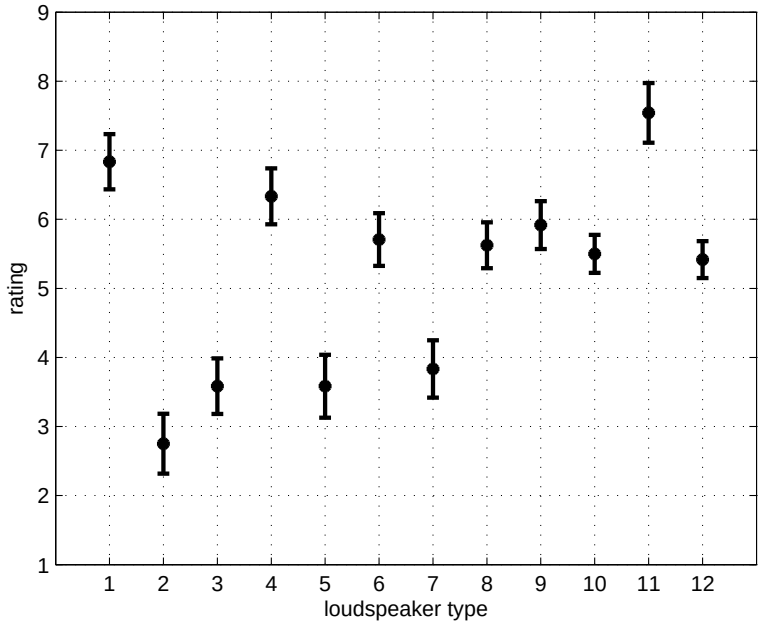


Figure 4.3: Rating of loudspeakers based only on visual appearance. Mean ratings and $\pm$ 1 standard error averaged for all subjects. The order of loudspeakers is the same as in figure 4.2 from top left to bottom right. A rating equal to 1 corresponds to low quality and 9 to high quality.

whether the loudspeakers could fit in a shelve, etc).

In manuscript E the attempt was to simplify the design and eliminate some of the factors that unnecessarily made the previous experiments more complex. The outcome of Exp5 indicates that simple experiments in terms of both stimuli and design might be a more appropriate approach. Easily differentiated stimuli and a design that creates unique stimuli combinations for each subject reduce potential sources of bias and confusion to the subjects. On the other hand, the Latin Squares design does not examine the interaction between modalities.

Audio dominates the audiovisual perception for this specific product under the setup used. Loudspeakers are a particular product choice as their main purpose is to reproduce sound, thus there exists an underlying semantic link between the specific product under test and the user's expectation of its properties. Results in this thesis show that subjects tend to focus on the sound fidelity and set aside style or aesthetic considerations. That being said, the influence of the degradations and the experimental question referring to sound in Exp1, Exp2, Exp3 and Exp4a, should be taken into account, and these results should not be generalized.

The stimuli presentation technique was shown to have little influence to the results. The presentation both for audio and visual stimuli was altered across experiments, however all main results were unaffected. Thus, for these specific experiments, with the selected design and stimuli, the reproduction mode for neither audio or visual is significantly affecting the results.

Bibliography

ITU-R Rec. BS.1286 (1997), Methods for the subjective assessment of audio systems with accompanying picture, International Telecommunications Union, Geneva, Switzerland.

Kohlrausch, A. and van de Par, S. (2005), "Audio-visual interaction in the context of multimedia applications", in J. Blauert (Ed.), "Communication Acoustics" (pp.109-138), Springer, Berlin, Germany.

Woszczyk, W., Bech, S., Hansen, V., (1995), "Interactions between audio-visual factors in a home theater system: Definition of subjective attributes", 99^{th} Audio Eng. Soc. Conv., New York, Preprint 4133:K-6.

Insko, B.E., (2003), "Measuring Presence: Subjective, Behavioral and Physiological Methods", in Riva, G., Davide, F. and Ijsselsteijn, W.A., (Eds.), "Being There: Concepts, effects and measurement of user presence in synthetic environments", Ios Press.

Thiede, T., Treurniet, W. C., Bitto, R., Schmidmer, C., Sporer, T., Beerends, J. G., Colomes, C. (2000), "PEAQ - The ITU Standard for Objective Measurement of Perceived Audio Quality", J. Audio Eng. Soc., Vol. 48, No. 1/2.

Rix, A. W., Hollier, M.P, Hekstra A.P, Beerends, J.G., (2002), "Perceptual Evaluation of Speech Quality (PESQ). The New ITU Standard for End-to-End Speech Quality Assessment Part I – Time-Delay Compensation", J. Audio Eng. Soc., Vol. 50, No. 10.

Takahashi, A., Hands, D. and Barriac, V., (2008), "Standardization Activities in the ITU for a QoE Assessment of IPTV", IEEE Communications Magazine, February 2008.

Puria, A., Chen, X. and Luthra, A., (1995), "A Distortion Measure for Blocking Artefacts in Images Based on Human Visual Sensitivity", IEEE Transactions on Image Processing, **4**(6), 1995.

Winkler, S., (2005), "Digital Video Quality: Vision Models and Metrics", Wiley, 2005.

ITU-T Rec. P.910 (2008), Subjective Video Quality Assessment Methods for Multimedia Applications. International Telecommunications Union, Geneva, Switzerland, 2008.

ITU-R Rec. BT.500-12 (2009), Methodology for the Subjective Assessment of the Quality of Television Pictures, International Telecommunications Union, Geneva, Switzerland, 2009.

ITU-T Rec. J.144 (2004), Objective Perceptual Video Quality Measurement Techniques for Digital Cable Television in the Presence of a Full Reference, International Telecommunications Union, Geneva, Switzerland, 2004.

ITU-R Rec. BT.1683 (2004), Objective Perceptual Video Quality Measurement Techniques for Standard Definition Digital Broadcast Television in the Presence of a Full Reference, International Telecommunications Union, Geneva, Switzerland, 2004.

Viollon, S., Lavandier, C. and Drake, C., (2002), "Influence of visual setting on sound ratings in an urban environment", Applied Acoust. 63, 493-511.

van Eijk R.L.J., Kohlrausch, A., Juola, J.F. and van de Par, S., (2008), "Audiovisual synchrony and temporal order judgments: Effects of experimental method and stimulus type", Percept. Psychophys., 70(6), 955-968.

Recanzone, G. H., (2003), "Auditory Influences on Visual Temporal Rate Perception", J. Neurophysiol. 89, 1078-1093.

Pick, H.L., Warren, D.H., Hay, J.C., (1969), "Sensory conflict in judgements of spatial direction", Percept. Psychophys., 6(4), 203-205.

Brown, J.M., Anderson, K.L., Fowler, C.A. and Carello, C., (1998), "Visual influences on auditory distance perception", J. Acoust. Soc. Amer., 104(3), 1798-1798.

MacDonald, J. and McGurk, H., (1978), "Visual influences on speech perception processes", Percept. Psychophys., 24(3), 253-257.

Alais, D., Morrone, C. and Burr, D., (2006), "Separate attentional resources for vision and audition", Proc. Biol. Sci., 273(1592), 1339-1345.

Brefczynski, J.A. and DeYoe, E.A, (1999), "A physiological correlate of the "spotlight" of visual attention", Nat. Neurosci. 2(4), 370-374.

Calvert, G.A, Campbell, R. and Brammer, M.J., (2000), "Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex", Curr. Biol. 10(11), 649-657.

Allport, D.A., Antonis, B. and Reynolds, P., (1972), "On the division of attention: a disproof of the single channel hypothesis", Q. J. Exp. Psychol. 24(2), 225-235.

Spence, C., Ranson, J. and Driver, J., (2000), "Cross-modal selective attention: on the difficulty of ignoring sounds at the locus of visual attention.", Percept. Psychophys. 62(2), 410-424.

Taylor, M.M., Lindsay, P.H. and Forbes, S.M., (1967), "Quantification of shared capacity processing in auditory and visual discrimination", Acta Psychol., 27, 223-229.

Massaro, D.W. and Warner, D.S, (1977), "Dividing attention between auditory and visual perception", Percept. Psychophys., 21, 569-574.

Stein, B.E. and Meredith, M.A, (1993), "The Merging of the Senses", Cambridge, MA: MIT Press.

Spence, C., (2007), "Audiovisual multisensory integration", Acoust. Sci. & Tech. 28, 2, 61-70.

Hollier, M.P. and Voelcker, R.M., (1997), "Towards a multi-modal perceptual model", BT Technology Journal, Vol. 14, No. 4, 163-172.

Molholm, S., Ritter, W., Javitt, D.C. and Foxe, J.J, (2004), "Multisensory visual-auditory object recognition in humans: A high-density electrical mapping study", Cerebral Cortex, 14, 452-465.

Laurienti, P.J., Kraft, R.A., Maldjian, J.A., Burdette, J.H., Wallace, M.T., (2004), "Semantic congruence is a critical factor in multisensory behavioural performance", Exp. Brain Res., 158, 405-414.

Vatakis, A. and Spence, C., (2008), "Evaluating the influence of the "unity assumption" on the temporal perception of realistic audiovisual stimuli", Acta Psychologica, 127(1), 1223.

Beerends J.G, and De Caluwe, F.E., (1999), "The influence of video quality on perceived audio quality and vice versa", J. Audio Eng. Soc., Vol. 47, No. 5, 355-362.

Hands, D.S., (2004), "A Basic Multimedia Quality Model", IEEE Transactions On Multimedia, 6(6), 806-816.

Hollier, M.P., Rimell, A.N., Hands, D.S. and Voelcker, R.M., (1999), "Multimodal perception", BT Technology Journal, 17(1), 35-46.

Zielinski, S.K., Rumsey, F., Bech, S., (2003), "Effects of Bandwidth Limitation on Audio Quality in Consumer Multichannel Audiovisual Delivery Systems", J. Audio Eng. Soc., Vol. 51, No. 6, 475-501.

Bech, S., Hansen, V., Woszczyk, W., (1995), "Interaction Between Audio-Visual Factors in a Home Theater System: Experimental Results", 99^{th} Audio Eng. Soc. Conv., 1995, October 6-9, NewYork.

ITU-R Rec. BS.1116 (1997), Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems, International Telecommunications Union, Geneva, Switzerland.

ITU-R Rec. BS.775-2 (2006), Multi-channel Stereophonic Sound System with or without Accompanying Picture", International Telecommunications Union, Geneva, Switzerland.

ITU-R Rec. BS.1534-1 (2003), Method for the subjective assessment of intermediate quality level of coding systems, International Telecommunications Union, Geneva, Switzerland.

Vatakis, A. and Spence, C., (2007), "Crossmodal binding: Evaluating the "unity assumption" using audiovisual speech stimuli", Percept. Psychophys., 69(5), 744-756.

ITU-T Rec. P.930 (1996), Principles of a reference impairment system for video, International Telecommunications Union, Geneva, Switzerland.

Blauert, J. and Jekosch, U. , (1997), "Sound-Quality Evaluation - A Multi-Layered Problem", Acta Acustica united with Acustica, Vol.83, No.5, 747-753.

Winkler, S., (1999), "Issues in Vision Modeling for Perceptual Video Quality Assessment", Signal Processing, vol. 78, no. 2, pp. 231-252.

ITU-T Rec. P.800 (1996), Methods for objective and subjective assessment of quality, International Telecommunications Union, Geneva, Switzerland.

Montgomery D., (2001), "Design and Analysis of Experiments", 5^{th}ed., Wiley, pp. 80.

5

Appendix

5.1 Statistical Analysis

As already discussed in the thesis and manuscripts, in principle ANOVA should not be applied to ordinal data, however the common viewpoint in the scientific community is that ANOVA can be applied to ordinal data (the F-test is robust to violations of the homogeneity of variances) (ITU-T Rec. P.800, 1996). The same is true for data that is not normally distributed. The major issue it the interpretation of the results. In the case of a significant difference, it is then valid to report that one group mean is higher or lower than another group mean - an ordinal statement. On the other hand making interval statements such as "group one is twice as much as the other group" should be avoided. The data from some of the experiments presented in this thesis were shown to be not normally distributed and all data were ordinal. This section presents the methods used to normalize the data, as well as methods for non-parametric analysis and a comparison to ANOVA.

5.1.1 Data normalization

The data from some experiments were shown to be not normally distributed. Furthermore, subjects might have different criteria and different approaches to the use of the rating scale. To overcome this problem and make the data more suitable for ANOVA, two approaches to normalizing the data described in (Viollon et al., 2002, 2002) and (ITU-R Rec. BS.1116, 1997) were tested. The data used for this example are from the audiovisual part of Exp4a. The same was done for Exp1 and Exp3 that were shown to exhibit a pattern of be non-normality similar to Exp4a distributed and where the application of normalization techniques resulted in similar outcomes to the ones presented in this section.

The deviation from normality of the data is examined here with the aid of normal probability plots (data are plotted against a theoretical normal distribution in such a way that the points should form an approximate straight line) plots of the residuals (residuals plotted against fitted values [1]), and the Kolmogorov-Smirnov test. Figures 5.1 and 5.2 show the data distribution before any attempt to normalize. The data is not normally distributed, but has an S shape indicating shorter than normal tails, i.e. less variance than expected. The plots of residuals against fitted values show a diamond shaped structure. This could indicate non constant variance (Montgomery, 2001). Figures 5.3, 5.4 and 5.5 show the variance for the levels of each factor. The

[1] Fitted values are actual observations minus the residuals

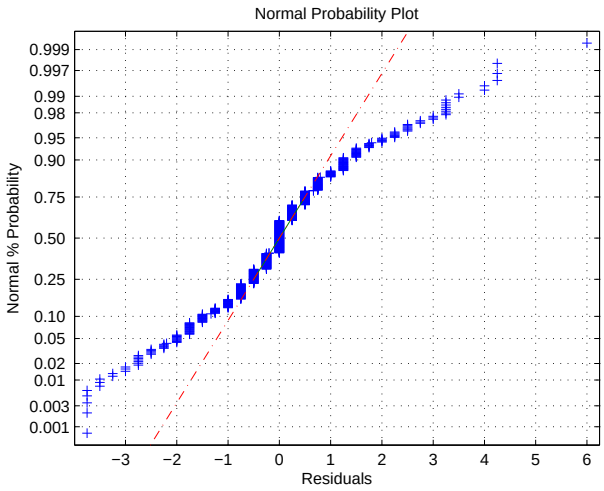


Figure 5.1: Normal probability plot for the non-normalized data of Exp4a. The dotted line shows a theoretical normal distribution.

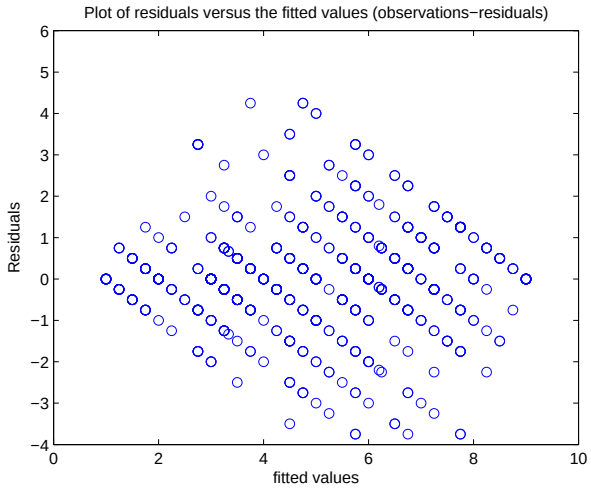


Figure 5.2: Residuals plot for the normalized data non-normalized data of Exp4a.

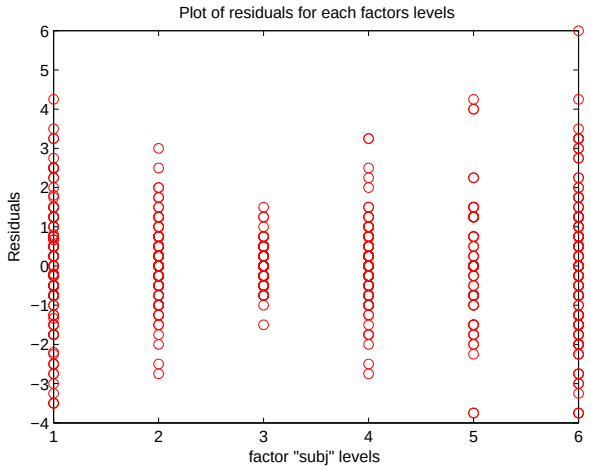


Figure 5.3: Residuals plot for factor *subjects* in the non-normalized data.

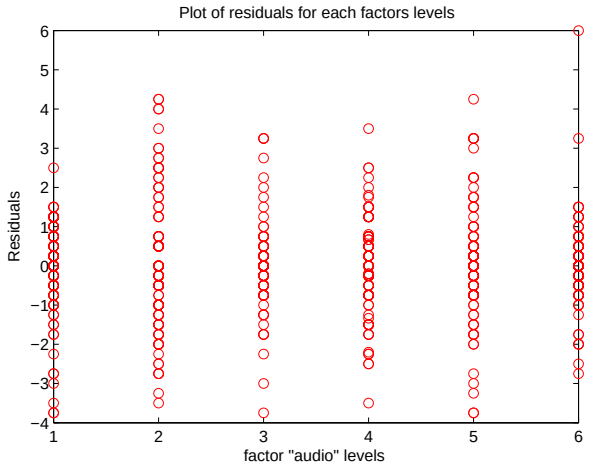


Figure 5.4: Residuals plot for factor *audio* in the non-normalized data.

figure for factor *subjects* shows some indication of inequality of variance.

Normalization according to (Viollon et al., 2002)

In order to eliminate the effect of the different usage of the rating scale by each subject, the data is normalized by subject (calculations performed for each subject).

$$X_{norm} = (X - \hat{X})/std$$

where:

X_{norm} : normalized data
X : initial data
$\hat{X}$: mean of all data
std : standard deviation of all data

The normal probability plot (figure 5.6) shows that the normalized data deviates less from a normal distribution. The residuals versus fitted values plot (figure 5.7) indicates a more evenly distributed variance. A Kolmogorov-Smirnov test shows that the normalized data is still not normally distributed.

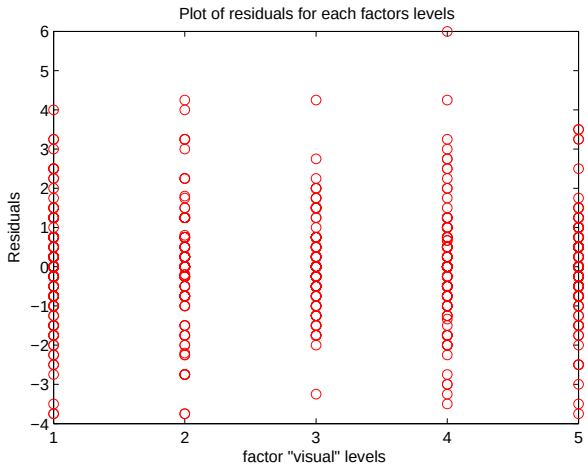


Figure 5.5: Residuals plot for factor *visual* in the non-normalized data.

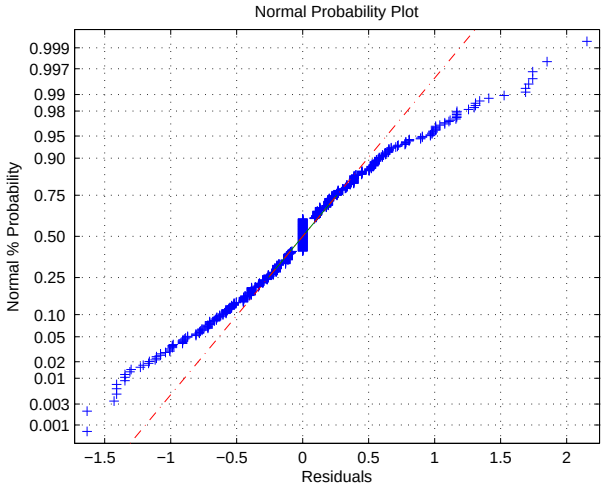


Figure 5.6: Normal probability plot for the normalized data according to the (Viollon et al., 2002) method.

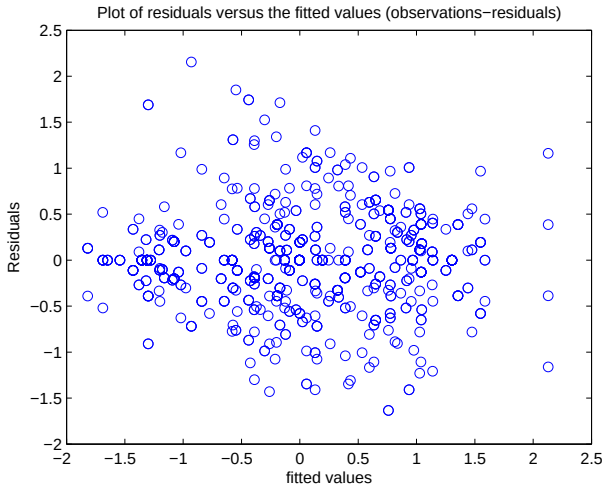


Figure 5.7: Residuals plot for the normalized data according to the (Viollon et al., 2002) method.

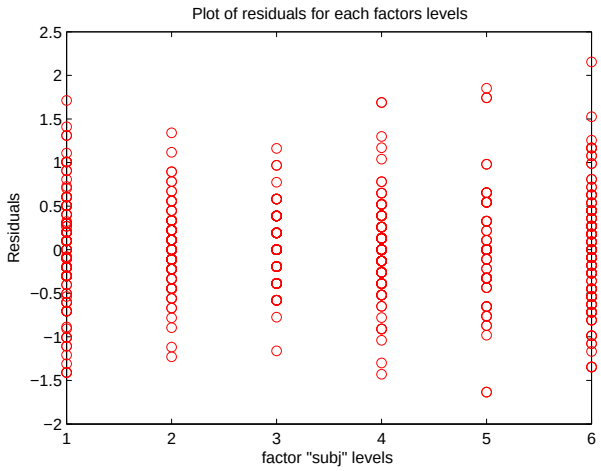


Figure 5.8: Residuals plot for factor *subjects* in the normalized data according to the (Viollon et al., 2002) method.

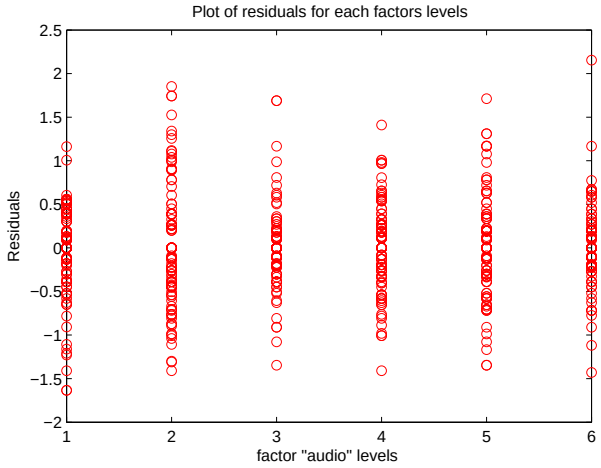


Figure 5.9: Residuals plot for factor *audio* in the normalized data according to the (Viollon et al., 2002) method.

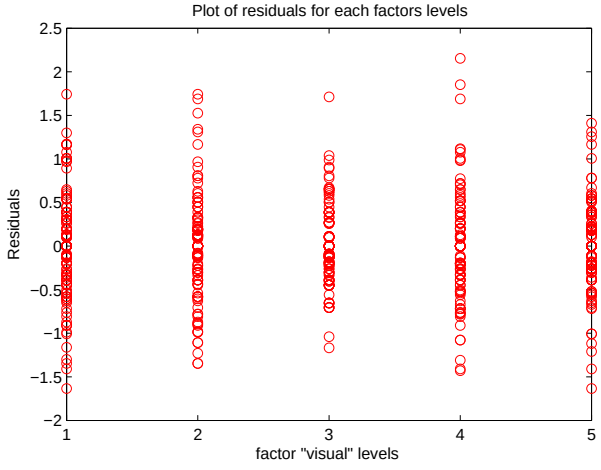


Figure 5.10: Residuals plot for factor *visual* in the normalized data according to the (Viollon et al., 2002) method.

Analysis of Variance

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
S	0.037	5	0.0074	0.02	0.9998
A	378.026	5	75.6051	215.47	0
V	1.26	4	0.3149	0.9	0.4652
S*A	101.024	25	4.041	11.52	0
S*V	6.243	20	0.3121	0.89	0.6009
A*V	8.067	20	0.4034	1.15	0.2943
S*A*V	29.416	100	0.2942	0.84	0.8612
Error	189.477	540	0.3509		
Total	714	719			

Constrained (Type III) sums of squares.

Figure 5.11: ANOVA for the normalized data according to the (Viollon et al., 2002) method.

Figures 5.8, 5.9 and 5.10 show the variance for the levels of each factor. The figures show no indication of inequality of variance.

The ANOVA for the normalized data according to the (Viollon et al., 2002) method is shown in figure 5.11. The ANOVA results are comparable to the ANOVA of the non-normalized data, with the only difference in factor *subjects* which is now statistically non-significant. For all other terms the conclusions are the same.

Normalization according to (ITU-R Rec. BS.1116, 1997)

The normalization according to (ITU-R Rec. BS.1116, 1997) is performed as follows:

$$Z_i = \frac{(x_i - x_{si})}{s_{si}} * s_s + x_s$$

where:

Z_i is the normalized result
x_i is the score of subject i
x_{si} is the mean score for subject i in session s
x_s is the mean score of all subjects in session s
s_s is the standard deviation for all subjects in session s and
s_{si} is the standard deviation for subject i in session s.

The conclusions that can be drawn in general are that the results are similar to those of the (Viollon et al., 2002) normalization. The normal probability plot (figure 5.12) shows that the residuals still deviate form a normal distribution but less than in the non-normalized case or the normalization according to Viollon et al., 2002. The plot of residuals versus the fitted values in figure 5.13 shows a more homogeneous variance. However, the Kolmogorov-Smirnov test shows that the normalized data is not normally distributed.

The ANOVA for the normalized data is shown in figure 5.17. The ANOVA results are comparable to the ANOVA of the non-normalized data, with the only difference in factor *subjects* which is now statistically non-significant. For all other terms the conclusions are the same.

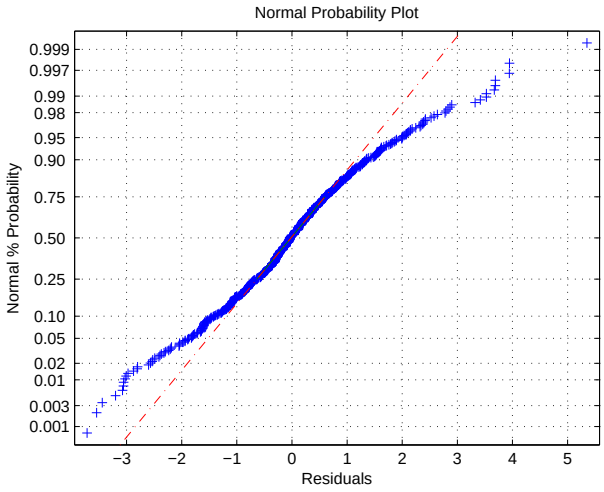


Figure 5.12: Normal probability plot for the normalized data according to the ITU method.

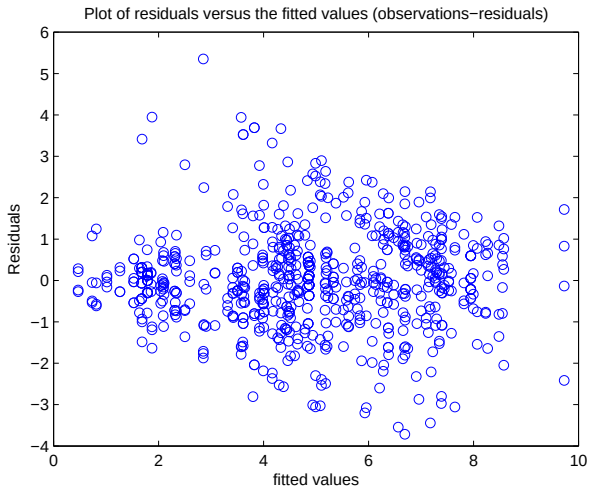


Figure 5.13: Residuals plot for the normalized data according to the ITU method.

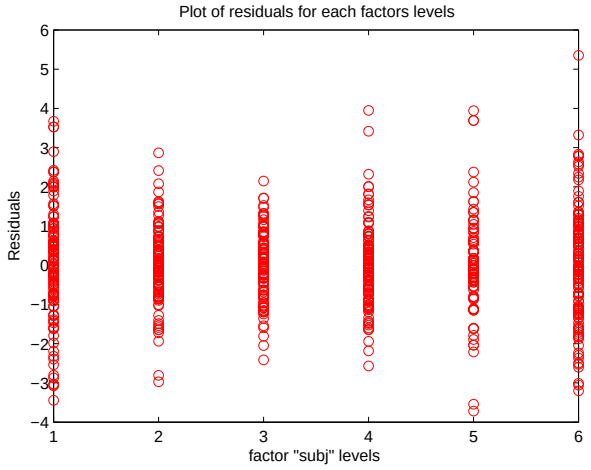


Figure 5.14: Residuals plot for factor *subjects* in the normalized data according to the ITU method.

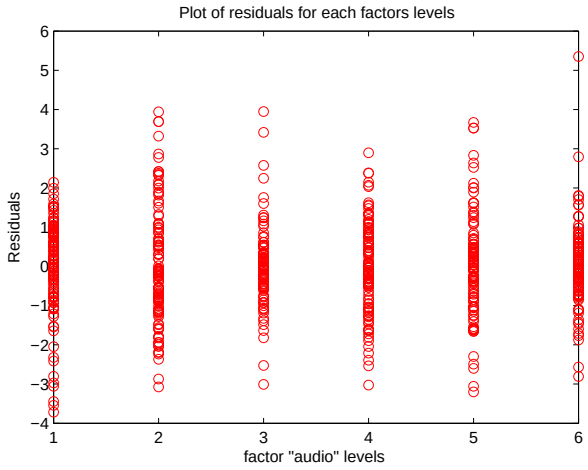


Figure 5.15: Residuals plot for factor *audio* in the normalized data according to the ITU method.

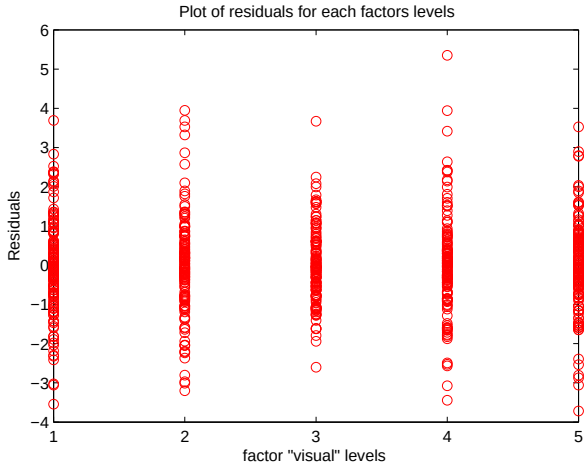


Figure 5.16: Residuals plot for factor *visual* in the normalized data according to the ITU method.

Analysis of Variance

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
S	0.18	5	0.037	0.02	0.9998
A	2096.77	5	419.355	232.86	0
V	7.4	4	1.849	1.03	0.3929
S*A	557.25	25	22.29	12.38	0
S*V	32.51	20	1.626	0.9	0.584
A*V	43.88	20	2.194	1.22	0.2326
S*A*V	159.4	100	1.594	0.89	0.772
Error	972.5	540	1.801		
Total	3872.45	719			

Constrained (Type III) sums of squares.

Figure 5.17: ANOVA for the normalized data according to the ITU method.

5.1.2 Non-parametric analysis

A range of comparisons between parametric (ANOVA) and non-parametric tests (Kruskal-Wallis, Friedman) is presented here. The reason for looking into non-parametric tests is that the data from all experiments are ordinal data instead of interval or ratio data. Strictly speaking, ANOVA requires either interval or ratio data. Here we compare the results of the same data with ANOVA and non-parametric tests. These tests do not assume a certain underlying distribution of the data and allow for ordinal data. All the analysis is based on the non-normalized data of Exp3.

There are small differences among the results of the parametric and non-parametric analysis, but the conclusions are the same with the statistical significance of all factors remaining the same. Furthermore, ANOVA has obvious advantages in that it includes analysis on the interactions, which is important for multimodal experiments. Note that in the Friedman test ANOVA table there appears an *interaction* term. This term shows the variability due to the interaction between rows and columns (if there are repetitions) , where *Columns* represent changes in factor A and *rows* represent changes in a blocking factor B. However, the analysis does not test for row effects or interaction effects.

The following tables present data of non-parametric tests, followed by the respective parametric test. Since the non-parametric tests presented here allow for the analysis of either 1 or 2 factors (Kruskal-Wallis and Friedman tests respectively), there are multiple tests of the same data set.

Friedman's ANOVA Table

Source	SS	df	MS	Chi-sq	Prob>Chi-sq
Columns	682372.1	5	136474.4	399.61	0
Interaction	5069.1	20	253.5		
Error	533483.3	690	773.2		
Total	1220924.5	719			

Test for column effects after row effects are removed

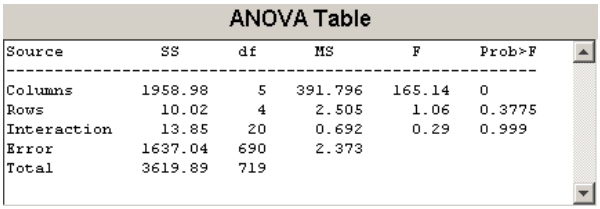
Figure 5.18: Friedman table for the AV data. Both factor *audio* and factor *visual* are included in the analysis but factor *subjects* is ignored. *Columns* refers to factor *audio*.

Friedman's ANOVA Table

Source	SS	df	MS	Chi-sq	Prob>Chi-sq
Columns	4499	4	1124.76	3.92	0.4163
Interaction	9172.7	20	458.64		
Error	804774.8	690	1166.34		
Total	818446.5	719			

Test for column effects after row effects are removed

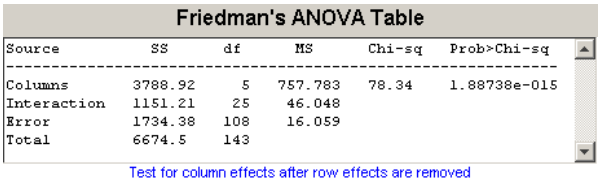
Figure 5.19: Friedman table for the AV data. Both factor *audio* and factor *visual* are included in the analysis but factor *subjects* is ignored. *Columns* refers to factor *visual*.



ANOVA Table

Source	SS	df	MS	F	Prob>F
Columns	1958.98	5	391.796	165.14	0
Rows	10.02	4	2.505	1.06	0.3775
Interaction	13.85	20	0.692	0.29	0.999
Error	1637.04	690	2.373		
Total	3619.89	719			

Figure 5.20: 2-way ANOVA table for the AV data. Both factor *audio* and factor *visual* are included in the analysis but factor *subjects* is ignored. *Columns* refers to factor *audio* and *rows* to factor *visual*.

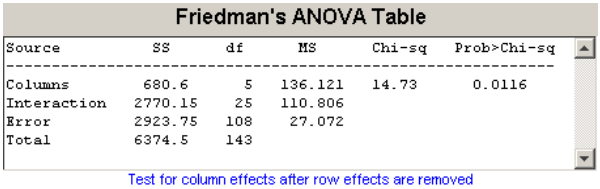


Friedman's ANOVA Table

Source	SS	df	MS	Chi-sq	Prob>Chi-sq
Columns	3788.92	5	757.783	78.34	1.88738e-015
Interaction	1151.21	25	46.048		
Error	1734.38	108	16.059		
Total	6674.5	143			

Test for column effects after row effects are removed

Figure 5.21: Friedman table for the A-only data. Both factor *audio* and factor *subjects* are included in the analysis. *Columns* refers to factor *audio*.

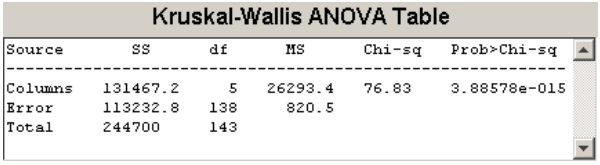


Friedman's ANOVA Table

Source	SS	df	MS	Chi-sq	Prob>Chi-sq
Columns	680.6	5	136.121	14.73	0.0116
Interaction	2770.15	25	110.806		
Error	2923.75	108	27.072		
Total	6374.5	143			

Test for column effects after row effects are removed

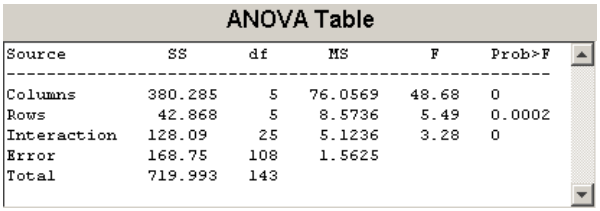
Figure 5.22: Friedman table for the A-only data. Both factor *audio* and factor *subjects* are included in the analysis. *Columns* refers to factor *subjects*.



Kruskal-Wallis ANOVA Table

Source	SS	df	MS	Chi-sq	Prob>Chi-sq
Columns	131467.2	5	26293.4	76.83	3.88578e-015
Error	113232.8	138	820.5		
Total	244700	143			

Figure 5.23: Kruskal-Wallis table for the A-only data. Since the test can handle only 1 factor, factor *subjects* is ignored.



ANOVA Table

Source	SS	df	MS	F	Prob>F
Columns	380.285	5	76.0569	48.68	0
Rows	42.868	5	8.5736	5.49	0.0002
Interaction	128.09	25	5.1236	3.28	0
Error	168.75	108	1.5625		
Total	719.993	143			

Figure 5.24: 2-way ANOVA table for the A-only data. Both factor *audio* and factor *subjects* are included in the analysis. *Columns* refers to factor *audio* and *rows* refers to factor *subjects*.

ANOVA Table

Source	SS	df	MS	F	Prob>F
Columns	380.285	5	76.0569	30.9	0
Error	339.708	138	2.4617		
Total	719.993	143			

Figure 5.25: 1-way ANOVA table for the A-only data. Since the test can handle only 1 factor, factor *subjects* is ignored.

Friedman's ANOVA Table

Source	SS	df	MS	Chi-sq	Prob>Chi-sq
Columns	707.96	4	176.99	21.55	0.0002
Interaction	2712.29	20	135.615		
Error	325.25	90	3.614		
Total	3745.5	119			

Test for column effects after row effects are removed

Figure 5.26: Friedman table for the V-only data. Both factor *visual* and factor *subjects* are included in the analysis. *Columns* refers to factor *visual*.

Friedman's ANOVA Table

Source	SS	df	MS	Chi-sq	Prob>Chi-sq
Columns	2901.2	5	580.24	62.35	3.96772e-012
Interaction	2049.8	20	102.49		
Error	400	90	4.444		
Total	5351	119			

Test for column effects after row effects are removed

Figure 5.27: Friedman table for the V-only data. Both factor *visual* and factor *subjects* are included in the analysis. *Columns* refers to factor *subjects*.

Kruskal-Wallis ANOVA Table

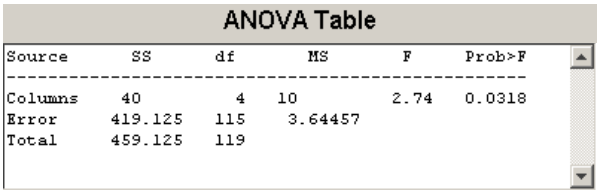
Source	SS	df	MS	Chi-sq	Prob>Chi-sq
Columns	16699.2	4	4174.81	14.22	0.0066
Error	123070.3	115	1070.18		
Total	139769.5	119			

Figure 5.28: Kruskal-Wallis table for the V-only data. Since the test can handle only 1 factor, factor *subjects* is ignored.

ANOVA Table

Source	SS	df	MS	F	Prob>F
Columns	40	4	10	42.35	0
Rows	206.375	5	41.275	174.81	0
Interaction	191.5	20	9.575	40.55	0
Error	21.25	90	0.2361		
Total	459.125	119			

Figure 5.29: 2-way ANOVA table for the V-only data. Both factor *visual* and factor *subjects* are included in the analysis. *Columns* refers to factor *visual* and *rows* to factor *subjects*.



ANOVA Table

Source	SS	df	MS	F	Prob>F
Columns	40	4	10	2.74	0.0318
Error	419.125	115	3.64457		
Total	459.125	119			

Figure 5.30: 1-way ANOVA table for the V-only data. Since the test can handle only 1 factor, factor *subjects* is ignored.